UNIVERSITY OF TARTU

DEPARTMENT OF ENGLISH STUDIES

A CORPUS BASED STUDY OF FORMULAIC LANGUAGE USE BY NATIVE AND NON-NATIVE SPEAKERS

BA thesis

ANDREAS PIIRI

SUPERVISOR: Jane Klavan, PhD

TARTU

2020

ABSTRACT

Although language makes use of formulaic patterns, knowing and using these formulaic patterns of words can prove to be quite difficult for non-native speakers of English. Since knowledge on formulaic language can both improve learners' comprehension and production of the language, it might be important for learners to familiarize themselves with these formulaic patterns. The aims of this thesis are to analyze whether or not non-native speakers use formulaic language more in their writing or in their speech as well as to compare the formulaic language use between native and non-native spoken language. Therefore a corpus-based analysis was conducted, which utilized the *Tartu Corpus of Estonian Learner English* (TCELE) for the written corpus, *Loucain International Database of Spoken English Interlanguage* (LINDSEI-EST) for the non-native spoken corpus, and *Michigan Corpus of Academic Spoken English* (MICASE) for the native spoken corpus.

The thesis begins with an introduction, which both gives an overview of the motivation for this paper as well as a summary for the following chapters. The literature review part of the thesis gives a definition for formulaic sequences, explains formulas and their usefulness, gives an explanation of n-grams in the study of formulas, and discusses the use of formulas in non-native language. Overviews of previous studies are given. The empirical part of the thesis at hand introduces the methodology, which consists of the used corpora and a two part analysis of the corpus data in the context of formulaic language, followed by a short analysis of the results. A longer, more detailed analysis of the results is given in the discussion part of the thesis. The thesis ends with a conclusion.

TABLE OF CONTENTS

INTRODUCTION

Language seems to make use of formulaic patterns of words or formulas. These formulas are semi-preconstructed even though they seem like they could be analyzed into segments. Understanding and using these formulas can prove difficult. However, knowing these formulas can improve the user's comprehension and production of the language. These formulas seem to be common in written language but even more common in spoken language (Leech 2002). Furthermore, it is thought that while native speakers tend to use a wide variety of different formulas, non-native speakers tend to use a more limited amount of formulas which they often over-use. This begs two questions. The first one being whether or not formulas are more often used in learners' written or spoken language. The second question would be whether non-native speakers use a more limited variety of formulas which they overuse? The study was conducted using the Michigan Corpus of Academic Spoken English (MICASE) for native spoken data, the Estonian subcorpus of Louvain International Database of Spoken English Interlanguage (LINDSEI-EST) for non-native spoken data, and Tartu Corpus of Estonian Learner English (TCELE) for non-native written data.

The first part of this thesis deals with what formulas are as well as why they are useful. This part will look at previous studies concerning formulaic language, as well as describe and analyze formulas. Because this study uses n-grams to determine the use of formulas by different speakers, this concept also needs to be defined and explained.

The second part of the thesis focuses on learner corpora. Since the study relies heavily on corpus data, it is important to define the term corpus and discuss the different types of corpora. Since learner corpora is a relatively new development it will also need to be defined. It is also beneficial to analyze the importance of corpora and what sort of information can be

extracted from them. Since analyzing just one type of corpus might not provide enough information, multiple corpora have to be analyzed to draw any conclusions from a study.

When the right corpora and analysis tools have been determined, a methodology has to be developed. This is what the third chapter of this thesis deals with. This chapter gives more details about the used corpora with metadata for both the written and spoken corpora. Information about the participants, length, and type of corpora is presented, as well as how the corpora were compiled. The second chapter then moves on to the first study, the aim of which is to determine whether learners of English use formulaic language more in written or spoken language. Here, the tools, the corpora, the process of cleaning the files, and the parameters are described. After the first study, the second analysis is discussed. This part focuses on the comparison between the usage of formulaic language between native and non-native speakers. Again, the tools, terminology, and corpora are described.

The third part of the thesis focuses on the results of the studies. Here, the results are displayed as well as analyzed. The second part of the third chapter will focus on the study concerned with the frequency of formulaic language usage in written and spoken language while the third part will focus on the comparison of formulaic language usage between native and non-native speakers.

The results are further analyzed in the discussion part of the thesis, where the implications of this study are discussed. In addition, options for further studies as well as which types of materials should be used for the further studies are discussed.

1. Formulas in native and non-native language

1.1 What are formulas and why are they useful?

According to Ellis et al. (2008), language uses many formulaic patterns of words, also known as formulas. In the research done by Sinclair (1991), it was said that a user of a

language has a large number of semi-preconstructed phrases available to them, these semi-preconstructed phrases may seem like they can be analyzed into segments but are single units.

According to Wray (2000: 465) a formulaic sequence is 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.

Formulaic sequences have been the focus of many studies in applied linguistics (e.g. Ellis et al. 2008; Granger & Meunier 2008). These formulaic sequences are considered to be central in both idiomatic and fluent use of language as well as in language acquisition (Ellis 1996). It seems that when learners' proficiency progresses, they start to prefer more open grammatical constructions and not just memorised, high functional-utility formulaic phrases. Therefore, formulaic sequences are thought to be psycholinguistically real (O'Donnell et al. 2013). The analysis of academic corpora seems to indicate that academic discourse contains common lexical bundles, collocations and formulaic sequences, and idioms at a high frequency (Ellis 1996; Ellis et al. 2008). Furthermore, previous comparisons between written and spoken corpora seem to indicate that these formulas are not only an important part of written language but an even more frequent occurrence in spoken language (Leech 2002). This is because speech is constructed in real-time, which relies on long-term memory and not calculation, therefore speakers rely on formulas more in spoken language (Bresnan 1999, Kuiper 1996).

According to the understanding of Grigaliūniene and Jukneviciene (2011: 14) the three main criteria of formulaic sequences are as follows: "they consist of more than one word, they recur in the corpus and they represent a cline of idiomaticity due to their internal semantic restrictions or, externally, pragmatic functions in the context."

Thus, there is a distinction between formulaic sequences and lexical bundles or clusters, for example. While formulaic sequences, according to Grigaliūniene and Jukneviciene (2011: 14) are required to have idiomaticity, lexical bundles and clusters do not. However, formulaic sequences can still come up in frequency based cluster lists (assuming that they fit the parameters given), but lexical bundles without idiomaticity or pragmatic meaning is not a formulaic sequence. For the purposes of this thesis paper, the term lexical bundle will be used, meaning that the bundles do not have to have distinct meaning or idiomaticity.

Knowing idioms, collocations and lexical bundles is important for learners not only for comprehension but also for nativelike production of the language (Ellis et al. 2008). However, even advanced learners can often face difficulties with these collocations and one common reason is the restrictions of the learners' first language influencing their ability to comprehend these formulas (Nesselhauf 2003). According to Ellis et al. (2008), research has shown that language processing makes use of formulaicity and collocation. In previous studies, it was found that the participants of a study considered idiomatic expressions such as *kick the bucket* to be meaningful much faster than non idiomatic expressions (Ellis et al. 2008: 376). Therefore, the prevalent nature of formulaic language is important for both learners and teachers of English.

Two ways that Ellis et al. (2008) analyzed these formulas was by analyzing the frequency and the mutual information (MI) of these formulaic patterns of words, the latter is a statistical measure used to determine how much do the words cohere and occur in a sentence (Ellis et al. 2008). The research done by Ellis et al. (2008) suggests that in both reading and recognition, and reading aloud, learners of English as a second language (ESL) seem to rely more on the frequency, while native speakers are more sensitive to MI.

1.2 N-grams in the study of formulas

By using n-grams we can predict the sequence of words in a sentence. For example, we intuitively know that very likely the next word in the sentence *Please turn your homework...* would be *in* or *over* but the words *refrigerator* or *the* are not very likely to follow (Jurafsky and Martin 2019). Probabilities are important for identifying words in, for example, speech recognition or handwriting recognition. These implications, however, go even further. Spellcheckers and machine translators also use the model of n-grams to identify mistakes or arrange a sentence correctly. These predictive models are called language models (LMs), the simplest of which is the n-gram, which assigns probabilities to sentences or sequences of words (Jurafsky and Martin 2019). Depending on the number of words we can call them bigrams, which consist of two words, trigrams consisting of three words, and so on.

One way to estimate the probability of word sequences is to take a large corpus and see how many times a sequence is followed by a specific word. This is called a relative frequency count. With a sufficiently large corpus, we can analyze these counts and estimate the probability (Jurafsky and Martin 2019). According to Ellis et al. (2008), natural language makes use of recurrent formulaic patterns of words. By using n-gram analysis we can, for example, analyze how much learners use these formulas in their writing or in their speech. Although we can identify the frequency of these clusters, this does not necessarily mean that those high frequency n-grams have meaningful, distinctive functions (Ellis et al. 2008). When analyzing a corpus for n-grams, it is important to set some parameters such as n-gram length, which is the length of the cluster, the frequency band, which is the minimum number of times the cluster has to occur within one text, and also range, which sets the minimum number of texts in which the n-gram has to occur to be reflected in the results.

According to O'Donnell et al. (2013), formulaic language should be objectively defined in terms of measurable operationalizations. In order to do this, n-grams can be used. The study by O'Donnell et al. (2013) examined formulaic language in learners through n-grams. They found that both frequency-defined and MI-defined formulas are more frequent in advanced writing at expert and graduate levels.

1.3 The use of formulas in non-native language

While previous experiments have shown native speakers' sensitivity to formulaic language, there is also considerable interest in formulaic language in second language acquisition and comparing native speakers' formulaic language usage with non-native formulaic language usage (Ellis et al. 2008: 377). In a three-part experiment by Ellis et al. (2008), it was shown that while native speakers' processing of the language is affected by the MI score, non-native learners' language processing is mostly affected by the frequency of the expression. Due to non-native speakers encountering high-frequency formulaic clusters more often, they are more familiar with them and will therefore use them preferentially. Non-native learners also have a lower input of language which means that they are much more affected by the frequency of the formulaic clusters rather than their MI scores.

It seems that learners progress from using memorized, high-frequency formulaic phrases to more open grammatical constructions (Ellis 2003). According to Grigaliūniene and Jukneviciene (2011: 14) the tendencies to use certain formulaic sequences are a part of the mastery of the language. With the emergence of corpus research the understanding of language ability was no longer only limited to just singular words but also other multi-word, fixed expressions (Grigaliūniene and Jukneviciene 2011: 14).

Paquot and Granger (2012) focused on studies which were explicit in their methodologies which allowed for comparisons to be made. Paquot and Granger (2012) made a distinction between co-occurrence and recurrence, pointing out that these evoke different competencies.

Co-occurrence, according to Paquot and Granger (2012), is a pattern that commonly consists of two lexical items, that can be contiguous. They go on saying that in traditional phraseology, these patterns are called (restricted) collocations. While a new frequency-based approach to phraseology uses the term collocation differently, Paquot and Granger (2012) used the term collocation in the traditional sense and the term co-occurrence to refer to statistically defined combinations, which are not as restricted. The main findings that Paquot and Granger (2012) came across concerning co-occurrence is that learners underuse, overuse, and misuse co-occurring combinations. While learners underuse some collocations, they also often overuse highly frequent collocations. Paquot and Granger (2012) point out that this might occur because learners feel comfortable using these collocations since they feel confident that they work.

Recurrence, according to Paquot and Granger (2012) is a string of contiguous words with a certain length, for example, bigrams which consist of two words and trigrams which consist of three words. They referred to these as lexical bundles, which can be grammatically complete string but they can also be incomplete. Overall they found that learners seemed to be more reliant on these lexical bundles, however, again repetition was prevalent and this might also be the consequence of learners copying the writing prompts of tasks.

2. Learner corpora

2.1 Native and learner corpora as a tool

Although the term "corpus" currently refers to a collection of written or spoken texts on a computer database (McCarthy 2004: 1), before the wide usage of computers it referred to a body of words. Analyzing these databases allows us to deduce various information, depending on the purpose of the corpus. Not only can we find out the frequency of words and phrases, but we can also use corpora to analyze the differences between spoken and written language and also the language usage of native speakers and learners. Digital corpora also allow us to search through and scan the collection of texts quickly (McEnery and Hardie 2011: 2) and give us an insight into how language is used in context (McCarthy 2004: 1).

Analyzing both the corpora of native speakers and learners, important information can be extracted on the most common mistakes of English learners and on the language usage of native speakers. Analyzing just one type of corpus, such as a corpus of native speakers, however, might not provide all the necessary information (Granger 2003: 534). The idea of a "learner corpus" is a relatively recent phenomenon, which started in the early 1990s with academics starting to collect and analyze learner language (Granger 2003). The interest and collection of these corpora have grown (Cotos 2014: 203). However, the exact definition of "learner corpus" is accompanied by uncertainty. Nesselhauf (2003: 127) suggests that since the idea of learner corpus is a relatively new development, the description is not yet systematic enough. One of the definitions is "an electronic collection of authentic texts produced by foreign or second language learners" (Granger 2003).

Learner corpora can be used as a tool for designing materials and to refine classroom methodology (Granger 2003: 542). Since investigating data only from grammaticality judgment tasks or choice tasks does not provide information on what and how learners

produce language instinctively, analyzing data on learner language production can provide important insight into this matter. Furthermore, analyzing the data provided by corpora can prove useful for dictionary writers to highlight common problems. As Xiao and McEnery (2010: 365) point out, dictionary writers used to come up with simple examples, because it was believed that learners would have difficulties comprehending authentic texts. This convention was, however, disrupted by the Collins Birmingham University International Language Database (COBUILD) project where authentic examples were used for learner dictionaries.

Learner corpora have also been used at the English department of Tartu University to research learner language. For example, Lemme Tammiste (2016), studied the use of adjective-noun, verb-noun and phrasal-verb-noun collocations in Estonian learner corpus of English. In addition Merli Kirsimäe (2016), wrote an MA thesis on a pragmatic analysis of a selection of interviews from an Estonian spoken mini-corpus of English as a lingua franca, and Anne Rahusaar (2019), wrote her master's thesis on the compilation of the spoken sub-corpus for the Tartu corpus of Estonian learner English. Furthermore, Aare Undo (2018) wrote his MA thesis on calculating the error percentage of an automated part-of-speech tagger when analyzing Estonian learner English, and Anna Daniel (2015) wrote her MA thesis on the use of adjectives and adverbs in Estonian and British student writing.

2.2 Different types of learner corpora

There is more than one type of learner corpus. Very commonly a distinction is made between general and specific, written and spoken, synchronic and longitudinal, mono-L1 and multi- L1 data. While general learner corpora could be used to show language usage in all contexts, specific learner corpora take into account a specific context or users (Gabrielatos

2005). Furthermore, in contrast to synchronic corpora, which consists of data from different learners at a specific, single period of time, longitudinal learner corpus looks at data from the same learners over a long period of time.

Since learner corpus is more contextualized and thorough it differs from regular error analysis. By only analyzing errors there could be significant omissions in the information about how learners use the language as pointed out by Ellis (1994: 64). Ellis (1994: 64) suggests that the only way to make generalizations about learners' language usage is to analyze both what the learner does correctly and where the learner is mistaken. Furthermore, the context in which words and phrases are used also has to be taken into account to see how aspects of language are used.

There are also different types of spoken learner corpora. While some corpora provide only the written transcription of speech, others can also include the audio recordings of the speech (Gilquin 2015). However, often the audio recordings are not accessible which means that only the transcription of the audio recording is used for the data of the corpus and is analyzed in the same way as data which is written from the start. Ballier and Martin (2013: 35) point out that a distinction can be made between three different types of spoken learner corpora. A mute spoken corpus consists of transcripts constructed on the basis of the audio recording, while the truly speaking corpora also provide access to the original audio recordings.

3. Methodology and data

3.1 The corpora used in the empirical study

In order to determine whether or not Estonian EFL speakers use more formulaic clusters in spoken language than they do in written language, two corpora were used: the

Tartu Corpus of Estonian Learner English (TCELE) and the Estonian subcorpus of Louvain International Database of Spoken English Interlanguage (LINDSEI-EST). The written corpus consists of 127 entrance essays, which were written in 2014 as one part of the entrance examination to the English Language and Literature BA program. The task given to the participants was to write a 200-word essay, which was based on an academic article about the future of the English language. This academic article will be referred to as the source text later on in the paper. The most important features of the corpus are: the length of the essays vary from 60 to 320 words, averaging at 193 words; all of the participants are Estonian citizens although their native language is not specified; the age of the participants range from 18 to 35, averaging at 19; out of the 127 participants, 88 are female and 39 are male; no reference tools were available to the participants; any mistakes and illegible words were left in during the process of typing up the corpus. (Tammiste 2016)

For the spoken language analysis, 17 transcribed interviews from the LINDSEI-EST corpus were used. The interviewing process was as follows. During the first part of the interview, the interviewees were given three topics from which they could choose one to spontaneously talk about for three to five minutes, no prior preparation time was given to the participants. After having spoken for the duration of three to five minutes, the interviewer would ask additional questions concerning the topic that was just discussed as well as questions about their hobbies and university life. During the second part of the interview, the interviewee had to retell a story based on four pictures, which they had to interpret. After having told the story, the interviewer asked additional questions (Rahusaar 2019).

The most important features of the LINDSEI-EST spoken corpus are: the 17 interviews amount to 224 minutes of speech and the average length of the interviews is 13 minutes, ranging from 8 - 17.4 minutes, there were 17 participants, 5 male and 12 female; 11

of the 17 interviews were given by the third-year students of English philology at the University of Tartu and 6 of them were given by the Master's program students. The native language of all the students was Estonian and they all gave a general overview of their English language background in a learner profile prior to the interview. All of the interviews were recorded and then transcribed manually by various people who have been involved in the project over the years. The total word count for the transcribed text produced by the interviewees in the LINDSEI-EST corpus is 21,066 words.

To compare the formulaic language usage in the spoken language of native and non-native speakers, the aforementioned LINDSEI-EST spoken corpus was used and the results were compared to the Michigan Corpus of Academic Spoken English (MICASE). MICASE (2002) is a collection of transcribed speech from almost 1.8 million words from the University of Michigan. It contains data from a range of different speech events and has been compiled by Rita Simpson-Vlach (project manager 1997 to 2006), John Swales (faculty advisor), Sarah Briggs (testing advisor). The MICASE corpus includes dialogues as well as monologues from 15 speech events. It includes speech from both the staff and students. The recordings vary from 19-178 minutes in length. It is also noted on the website that the speech is not necessarily just scientific discussion but includes jokes and explanations as well. (Simpson et al. 2019)

The MICASE corpus allows for the following parameters to be set: speech event type, academic division, academic discipline, participant level, interactivity rating, gender, age, academic role/position, native speaker status, first language. For the purposes of this thesis the academic division was set to humanities and arts, the participant level was set to junior graduate and junior undergraduate, the native speaker status was set to native speaker, American English, and native speaker, other English. All genders and ages were included.

The rest of the parameters were left unselected. The total word count with these parameters selected is 111,116 words.

3.2 Analysis 1: Formulaic clusters in spoken and written language of Estonian learners of English

The first aim of the study was to find out whether or not learners use formulaic clusters more in spoken language, as is suggested by research (Bieber, Johansson, Leech, Conrad & Finegan, 1999; Brazil 1995). Therefore, a quantitative method was applied. The source data was gathered from the TCELE and LINDSEI-EST corpora. To find out the frequency of formulaic clusters in the corpora, a software called AntConc 3.5.8 (2019) by Laurence Anthony was used. The clusters/n-gram tool was used to determine the frequency of formulaic clusters in both the spoken transcriptions and written text. However, to eliminate as much statistical noise as possible, the spoken transcriptions had to be cleaned up. Due to the impulsive nature of the spoken language, they often include pauses that are filled with words such as "umm, erm, etc." For the purposes of this research, these fillers were removed by using the find and replace function in the text editor software called Notepad++ version 7.8.6 (2020) by Don Ho. All of the syntax tags of transcriptions were also removed, this was done by using the "hide tags" function in AntConc 3.5.8 (2019). Furthermore, because the transcriptions of spoken language use symbols such as " . " to represent pauses, they were also removed - this was done by using the find and replace function in the text editing software Notepad++ (2020).

After the files of the corpora were cleaned up, the next step was to import them into the AntConc 3.5.8 software and run the analysis. Before that, however, some parameters had to be set. The first parameter to be determined was the minimum frequency of clusters,

meaning, the minimum number of times a cluster had to appear in all texts. Ellis et al. (2008) used a minimum frequency of 10.9 per million. However, because of the corpus used for this study was much smaller, a lower minimum frequency of 3 was used. The next parameter to be set was the n-gram length, which determines the minimum and maximum length of the clusters. According to Ellis and Vlach (2010), it is known that two-word phrases not only have a very high occurrence frequency but are also commonly subsumed in 3- or 4- word phrases. Therefore, two-word phrases were excluded from the analysis. Furthermore, although five-word phrases are relatively rare, they were still included in the final analysis in an attempt to be as thorough as possible. The last parameter that had to be set was the minimum range, which determines the minimum number of files a cluster has to occur in. Because the corpora used for this study are small and the written essays were all gathered into one file, the minimum range was set to 1.

After the texts were cleaned up, imported into AntConc 3.5.8, the parameters set, the automatic analysis could be run. After the results were obtained, they were exported from AntConc 3.5.8 and saved as a text file, which could then be copied into Google Sheets for easier analysis. For the sake of thoroughness, the results were then scanned manually to confirm that no anomalies interfered with the statistics. After the results of the first analysis were gathered, it was discovered that the results might have been inaccurate and unreliable due to the formulaic clusters being influenced by the source text which was used for the essays of the TCELE written corpus. This is likely to have have occurred because of the way the corpus was compiled. Out of the top 30 formulas 18 of them seems to have been influenced by the source text used in the entrance essays, as is demonstrated in Table 1. These results will be described in greater detail at the end of the third part of this thesis paper.

Table 1. Top 30 n-grams in the written essay files

#Total No. of N-Gram Tokens: 7974	Freq	
1	135	new standard of
2	131	of international english
3	127	a new standard
4	119	standard of international
5	118	new standard of international
6	114	standard of international english
7	113	new standard of international english
8	111	a new standard of
9	99	a new standard of international
10	66	international english will
11	57	of international english will
12	53	english will emerge
13	53	standard of international english will
14	50	international english will emerge
15	47	of international english will emerge
16	44	that a new
17	43	it would be
18	40	that a new standard
19	39	that a new standard of
20	38	positive and negative
21	36	i think that
22	32	on the other
23	30	more and more
24	29	on the other hand
25	29	the other hand
26	28	over the world
27	28	there will be
28	27	in the world
29	26	all over the
30	26	all over the world

Because 18 out of the top 30 formulas seem to be influenced by the clause “that a new standard of international English will emerge”, the results of this first analysis were not taken

into account. In future research comparing the frequency of formulaic clusters in spoken and written language of advanced learners of English, it might be important to scan through the results and remove the formulaic clusters which are influenced by the source text or to use a different corpus for the data. However, even by removing the source text effect, there is still a possibility that the task which was used to compile the corpus would still have an effect on the data. In this case, the formula *positive and negative* could occur because of the task effect — the learners were given the following instructions:

> "According to Cook, it is likely that a new standard of international English will emerge. What might be some of the consequences (both positive and negative) of this for English as well as other languages? Provide reasons for your opinion. (Write an answer of approximately 200 words on your answer sheet.)" (Daniel 2015: 56)

An interesting case can be made, however, for the cluster *I think that*. This will be expanded upon in the discussion part of the thesis.

3.3 Analysis 2: Formulaic language usage in native and non-native speakers of spoken English

Because the comparison between the written and spoken corpora did not provide sufficiently clear results it was decided that the next course of action would be to compare the usage of formulaic language between learners of English and native speakers. However, before the comparative analysis could be started it was necessary to have a set of lexical bundles. For this study, the Academic Formulas List (AFL) (Ellis and Vlach 2010) was used as a basis for the comparative analysis. The AFL is a list of the most common formulaic bundles in academic English. It is similar to the Academic Word List but rather than listing singular words, the AFL lists three to five-word sequences. This list was developed and compiled by Rita Simpson-Vlach and Nick C. Ellis from the University of Michigan. The lists have been separated into three sections. The first one is the Core AFL list, which consists

of both written and spoken language. This core AFL list consists of 207 entries. The second list consists of only spoken academic language and has 200 entries. The last list consists of only written academic language and also has 200 entries. The lists also provide a statistic called formula teaching worth (FTW), which is a measure of usefulness. This statistic should indicate which formulaic bundles should be learned. For this research only the spoken AFL was used (see the full list in Appendix 1). This choice was made based on the claims that formulas are more common in spoken language and because the written TCELE corpus seems to be heavily influenced by the source text used in the compilation of the corpus.

The results of LINDSEI-EST were compared to the MICASE (2002) corpus. However, prior to the empirical analysis, it was important to set some relevant parameters for the MICASE corpus so that the results would only reflect the formulaic language usage of native speakers and that the context would be similar to the LINDSEI-EST spoken corpus. Data from all genders and all ages was included, for the native speaker status parameter the "Native speaker, American English" and "Native speaker, Other English" were chosen. To make sure that the context of formulaic language usage was as similar as possible to that of the LINDSEI-EST corpus, the next parameter to be set was Academic Division. This was set to "Humanities and Arts". For the participant level parameter, junior graduate and junior undergraduate was selected.

First, the list of 200 entries from AFL was compared to the LINDSEI-EST corpus, the full list of comparisons is given in Appendix 1. This was done manually by compiling a spreadsheet that includes all of the formulaic bundles from the spoken AFL. The first step in the analysis was finding the raw frequencies. This was done by going through each of the 200 entries and by using the Find function in Google Sheets. Each formulaic bundle provided by the spoken AFL was searched for in the results sheet of the LINDSEI-EST corpus. The

LINDSEI-EST frequencies were then copied into a spreadsheet for further analysis. Another statistic that was added was rank. This is a statistic that shows how high on the list a specific sequence was. This was analyzed in case any patterns would emerge. There were also two entries in the spoken AFL which included the clause "university of Michigan", these two entries were ignored, due to them being irrelevant in the context of formulaic language usage within non-native and native speakers.

After the frequency of all 200 entries from the spoken AFL were found for LINDSEI-EST and added into the spreadsheet, the process was replicated for the MICASE corpus. As mentioned before, some parameters were set for the MICASE corpus. This was done to make sure that only the formulaic language usage of native speakers would be provided and also to make sure that the context of the formulaic language usage was as similar to that of the LINDSEI-EST corpus as possible. Again, each of the 200 entries was searched for in the MICASE corpus. Since the MICASE corpus functions like a search engine, rather than being a static list, all of the 200 entries were entered manually into the search bar. Once the results were returned, the "view results statistics" tab was opened to get a more thorough result. The frequencies of the formulaic bundles were then copied into the spreadsheet for the MICASE corpus.

The mathematical concept of relative frequency was used to find out the frequency per 10 000 words. This was done to make sure that the results were comparable across the different corpora. Relative frequency is a mathematical concept used when comparing two or more corpora. In order to do this, the absolute frequency has to be determined as well as the total number of tokens in a corpus. The calculation involves dividing the absolute frequency by the number of tokens in the corpus and multiplying that with the basis of normalization. (Brezina 2018: 43)

The top 10 of the most frequent formulaic clusters for native speakers were: *This is the* (freq per 10 000 = 2.96), *a kind of* (freq per 10 000 = 2.96), *you can see* (freq per 10 000 = 2.42), *and then you* (freq per 10 000 = 1.97), *you need to* (freq per 10 000 = 1.88), *it could be* (freq per 10 000 = 1.88), *this kind of* (freq per 10 000 = 1.79), *take a look* (freq per 10 000 = 1.79), *and this is* (freq per 10 000 = 1.61), *the kind of* (freq per 10 000 = 1.61).

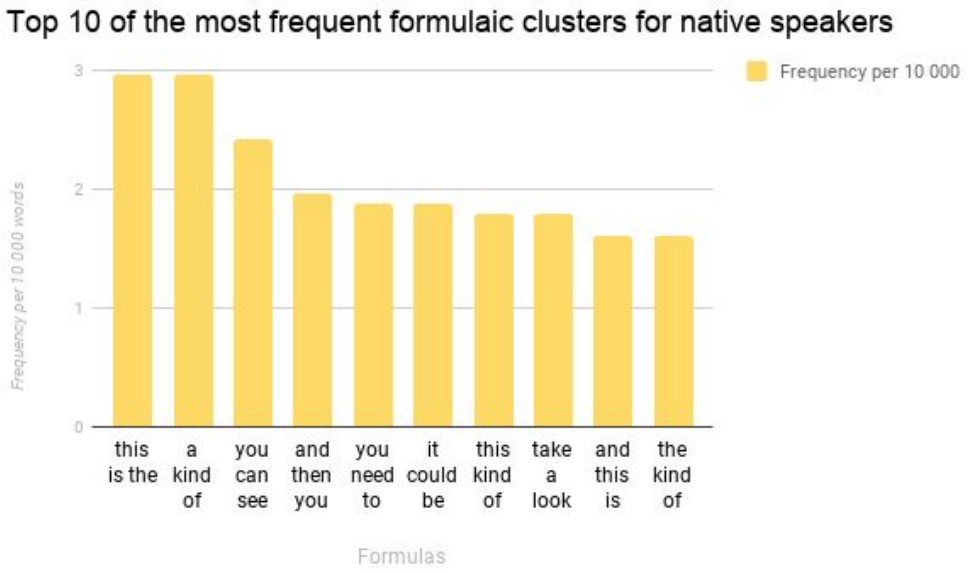

Figure 1. Top 10 of the most frequent formulaic clusters for native speakers.

The top 10 of the most frequent formulaic clusters for the non-native speakers were: *I wanted to* (freq per 10 000 = 2.74), *and then you* (freq per 10 000 = 2.29), *there was a* (freq per 10 000= 2.29), *yes yes yes* (freq per 10 000 = 1.83), *this kind of* (freq per 10 000 = 1.83), *you can see, look at the, the end of, no no no*, all with a frequency of 1.37 per 10 000.

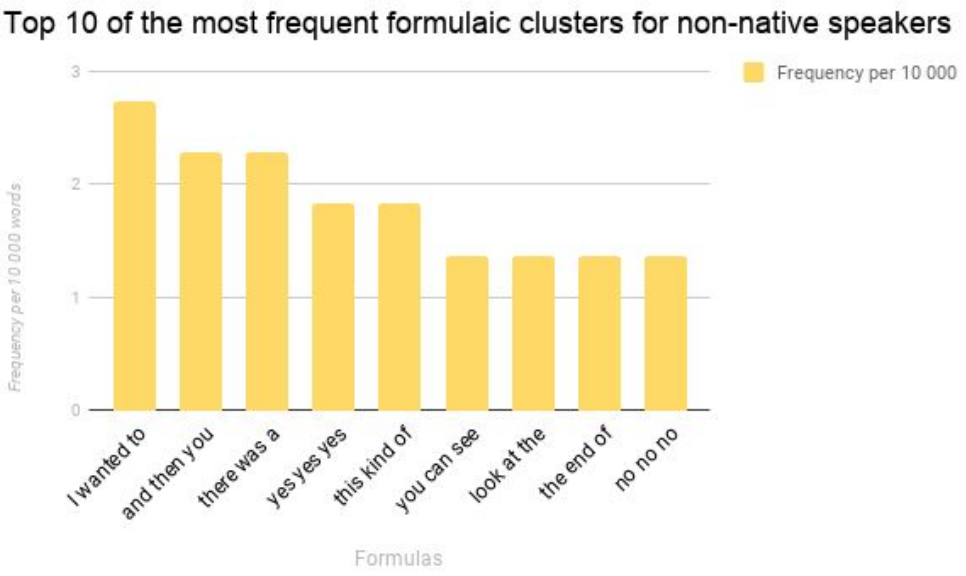

Figure 2. Top 10 of the most frequent formulaic clusters for non-native speakers

Only three of the 10 formulaic clusters were present in both the native and non-native top 10 most frequent formulas: *you can see, and then you, this kind of,* with *and then you* and *this kind of* occurring higher in the non-native top 10 most used formulas list. From the analyzed data, a clear pattern emerges. Out of the 198 formulaic clusters provided by the AFL, native speakers used 159 of them. In contrast, non-native speakers only used 12 of the 198 formulaic patterns. These results seem to match those of previous research (Paquot and Granger 2012), which suggests that the frequency of non-native speakers' formulaic language use is often less frequent when compared to the native speakers' use of formulaic language. However, no clear pattern emerged between the ranking of formulas between the AFL and non-native speakers' formulaic language use.

3.4 Discussion

While the analysis between non-native spoken and written language did not provide any clear results due to the interference of the source text used to compile the TCELE written

corpus, there are still a few interesting observations to make. A phenomenon concerning the lexical bundle *I think* occurred in this study. Out of 1,261 clusters that were retrieved using AntConc 3.5.8 from the TCELE written corpus, the cluster *I think that* was the 21st most frequent cluster with additional variations of *I think* also being present. The formulaic cluster *I think that* has been a point of interest in other studies as well. For example, the study by Grigaliūniene and Jukneviciene (2011) found that *I think* was one of the most frequent clusters in many corpora, both learner and native speaker.

According to Wierzbicka (2006: 37) the high frequency of *I think* may be connected to the Anglo respect for facts. Furthermore, *I think* emphasizes the distinction between what one knows from what one thinks. Whether or not non-native speakers use the cluster in a similar manner is difficult to tell without researching the context of the usage. Brown and Levinson (1987) pointed out that conversation markers are used for different reasons by native speakers and learners. Namely, while native speakers use conversational markers to express interpersonal functions which are used to be polite and indirect, learners tend to use them to express uncertainty and imprecision.

The case of the *I think* cluster is something that future studies could focus on. However, it is imperative that the context is taken into account and that the study would not just focus on frequency as the paper at hand did. In addition, the research paper at hand gathered data from the TCELE written corpus that was heavily influenced by the source text which was used in the compilation of the corpus. Future research, both quantitative and qualitative, should remove sequences influenced by the source text.

In the case of the second analysis, regarding the frequency of formulaic language usage between native and non-native speakers, the results are more definitive. As was suggested by the previous research by Paquot and Granger (2012), overall, learners may be

more reliant on lexical bundles which they often overuse. The variety in formulaic language usage in non-native speakers is less than that of native speakers'. As mentioned before, non-native speakers only used 12 of the 198 lexical bundles provided by the AFL. This is remarkably less than the 159 lexical bundles that native speakers used.

Furthermore, there is an interesting observation to be made about the frequency of lexical bundle usage in the case of non-native speakers. With the exception of the lexical bundle *you can see*, 11 of the lexical bundles used by non-native speakers were more frequently used compared to native speakers. This might indicate a case of what Hasselgren (1994) called a lexical teddy bear, which means that learners will often overuse certain basic words such as *very* instead of taking the risk of making an error with a less frequent word. This phenomena of the lexical teddy bear was also observed in the research concerning formulaic language and second language acquisition by Ellis (2012). There it was argued that the same concept could be applied to multi-word sequences and they called it the phrasal teddy bear.

As mentioned before the TCELE written corpus was influenced by the source text, however, the LINDSEI-EST spoken corpus is not completely free from task-effects either, which in turn would mean that the results of this thesis might not be conclusive either. Due to how the spoken corpora was compiled, it is possible that the participants could not express themselves completely freely. As mentioned in part 3.1, the first task involved participants being given a choice between three topics, which they had to spontaneously talk about, and after which they were asked additional questions concerning the topic. In the third task, the participants had to retell a story based on four pictures, after which, again, they were asked additional questions. While these tasks might make the interviewing process efficient, it might not provide the most natural speech patterns. Because of the fixed tasks, the data of the

corpora is going to be limited to data only regarding those tasks. In order to conduct future research with more accuracy, not only should the corpus of estonian spoken learner data be substantially larger, but the data should be more varied. For example, for the native speaker data, the MICASE corpus was compiled by recording and transcribing data from a range of different speech events. While compiling the spoken corpus using interviews with fixed tasks is a good start, including transcriptions from lectures and other speech events might improve the variety of data in the LINDSEI-EST corpus. This, in turn, would provide future research with not only more data, but more accurate data as well.

While the research at hand only focused on the frequency of formulaic language usage, the implications of formulaic language go further. For example, given enough data, the Academic Formulas List compiled by Ellis and Vlach (2010) could be improved upon. With enough data from a wide variety of participants, the list could change and the Formula Teaching Worth parameter could prove to be even more useful for both learners and teachers of English. Future research could focus on the usage of formulaic language between learners with different native languages or learners with different language proficiency levels. Assuming that the research would provide any significant results, the list could be broken down even further, for example, making separate lists for learners of different language proficiency, or even making separate lists for learners with different native language backgrounds. Again, this would require the data of the corpus to be as accurate as possible.

Conclusion

Language makes use of formulaic patterns of words. These formulaic patterns might seem as though they could be analyzed into segments, but they are, in fact, (semi)preconstructed. Although knowing and learning these formulas can prove useful for

both the language production and comprehension of a learner, using these formulas can be difficult for learners of English. These formulaic patterns are not only common in written language, but previous research has shown that they are at least as common or more common in spoken language as well. The thesis at hand analyzed the frequency of formulaic language usage in non-native written and spoken language and compared formulaic language usage between non-native and native spoken language.

The first part of the thesis explained what formulas are and stressed their usefulness. A definition for formulaic sequence was given. It was also explained that formulaic language has been a focus for many previous studies, with some examples given as well of these studies. It was pointed out that analyzing academic discourse indicates that lexical bundles, collocations, and idioms are used at a high frequency. Furthermore, according to some research, the usage of formulaic language can be even more common in spoken language. A brief explanation of why this might be was given. The importance of knowing these lexical bundles, collocations, and idioms was discussed. Various ways of analyzing data regarding lexical bundles were discussed and it was pointed out that for the purposes of this thesis, frequency would be analyzed. The phenomena of n-grams was explained, and the use of formulas in non-native language, specifically, was talked about in more detail.

Because the thesis at hand heavily utilizes corpus data, in the second part, the term corpus was explained and the emergence, and the benefits of learner corpora were discussed. First, using learner corpora as a tool was talked about. This included the origins of the term as well as the importance of analyzing both native and non-native corpora. Various kinds of learner corpora such as general vs specific, synchronic vs longitudinal, etc. were explained as well.

The third part of this thesis dealt with the methodology of gathering data from the corpora, and also included the results of the gathered data. First, the corpora which were used in the study were analyzed. The compilation process of both the TCELE written corpus and LINDSEI-EST spoken corpus was explained. In addition, the most important features of both the aforementioned corpora were discussed. Furthermore, since the second analysis required a native-speaker corpus, some details of the chosen MICASE corpus were pointed out.

Secondly, the methodology behind conducting the first analysis, as well as the results, were explained. This included gathering the data from the corpora, cleaning up the files, retrieving the data regarding lexical bundles using a software called AntCont 3.5.8, and the process of data analysis. Some of the most important findings were displayed. Although the first analysis did not provide the expected data, an interesting observation was made, which was discussed later in the thesis. Thirdly, the methodology behind the second analysis, as well as the results, were explained. This included explaining the use of the Academic Formulas List in the thesis at hand, the parameters set for retrieving accurate data from the MICASE corpus, the method for finding the raw frequencies for both the LINDSEI-EST and MICASE corpus, the concept of relative frequency and why it is useful for corpus research..

Finally, the results for both of the analysis' were discussed in more detail. This consisted of pointing out some observations regarding the lexical bundle *I think* and discussing implications as well as possibilities for future research, analyzing the results of the second analysis in which the phenomena of lexical teddy bear was pointed out, some possible flaws with the current corpus data and therefore the thesis were pointed out, and finally some further implications and possibilities for future studies were mentioned.

Researching formulaic language usage in non-native speakers can prove to be useful for both learners and teachers of the language. Although this particular thesis focused solely

on the frequency of formulaic language usage in non-native spoken and written language, and formulaic language usage in native and non-native spoken language, the implications of this type of research go further, some of which were mentioned in this paper. Having said that, a frequency based as used in the present study analysis should be considered as the first step in a more elaborate analysis. Further research should focus on qualitative analysis of a particular set of lexical bundles or formulaic sequences with a focus on the context of use.

List of references

Ballier, Nicholas. Martin, Philippe. 2013. Automatic Treatment and Analysis of Learner Corpus Data. Developing corpus interoperability for phonetic investigation of learner corpora. Amsterdam: John Benjamins Publishing Company.

Bieber, D. Johansson, S. Leech, G. Conrad S. & Finegan, E. (1999). Longman grammar of spoken and written English. Harlow, England: Pearson Education.

Brazil, David. 1995. A grammar of speech. Oxford: Oxford University Press.

Brezina, Vaclav. 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge University Press.

Bresnan, Joan. 1999. Linguistic theory at the turn of the century. Plenary presentation. Paper presented at the 12th World Congress of Applied Linguistics. Tokyo, Japan.

Brown, Penelope. Levinson, C. Stephen. *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Cotos, Elena. 2014 Enhancing writing pedagogy with learner corpus data. Volume: 26, 202-244.

Daniel, Anna. 2015. The use of adjectives and adverbs in Estonian and British student writing: a corpus comparison. Available at https://dspace.ut.ee/bitstream/handle/10062/47055/AnnaDaniel_MA.pdf?sequence=1&isAllowed=y, accessed May 25, 2020.

Ellis, C. Nick. 1994. *Implicit and Explicit Learning of Languages*. London: Academic Press.

Ellis, C. Nick. 1996. Sequencing in SLA: Phonological Memory, Chunking, and Points of Order. Studies in Second Language Acquisition, 18: 1, 91-126. Available at

https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/sequencing-in-sla/C8510F67FC125666556602B34E4F1EE4, accessed February 10, 2020.

Ellis, C. Nick. 2003. *Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure*. Oxford: Blackwell.

Ellis, C. Nick. 2012. Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual review of Applied linguistics*. 32, 17-44.

Ellis, C. Nick. Rita Simpson-Vlach. Garson Maynard. 2008. Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. Volume 42, 375-394.

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

Gabrielatos, Costas. 2005. Corpora and Language Teaching: Just a fling or wedding bells?. Available at https://files.eric.ed.gov/fulltext/EJ1068106.pdf, accessed February 13, 2020.

Gilquin, Gaëtanelle. 2015. From design to collection of learner corpora. Cambridge: Cambridge University Press.

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

Granger, Sylviane. Fanny, Meunier. 2008. Phraseology: An interdisciplinary perspective. Amsterdam: John Benjamins Publishing Company.

Grigaliūnienė, Jonė. Juknevičienė, Rita. 2011. Formulaic language, learner speech and the spoken corpus of learner English LINDSEI-LITH. Available at

https://etalpykla.lituanistikadb.lt/object/LT-LDB-0001:J.04~2011~1367179648554/J.04~2011~1367179648554.pdf , accessed May 12, 2020.

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

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

Jurafsky, Daniel. Martin, H. James. 2019. N-gram Language Models. Available at https://web.stanford.edu/~jurafsky/slp3/3.pdf, accessed January 30, 2020.

Kirsimäe, Merli (2016). Pragmatic analysis of a selection of interviews from an Estonian spoken mini-corpus of English as a lingua franca. Available at https://dspace.ut.ee/bitstream/handle/10062/63477/Park_Rannar_BA_Thesis.pdf?sequence=1&isAllowed=y, accessed May 24, 2020.

Kuiper, Koenraad. 1996. Smooth talkers: The linguistic performance of auctioneers and sportscasters.

Laurence, Anthony. 2019. AntConc (Version 3.5.8). [MacOS]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Leech Geoffrey 2002. Grammars of Spoken English: New Outcomes of Corpus-Oriented Research. *Language Learning.* 50: 4, 675-724.

McCarthy, Michael. 2004. Touchstone: From Corpus to Course Book. Cambridge: Cambridge University Press.

McEnery, Tony, Andrew, Hardie. 2011. Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.

Nesselhauf, Nadja. 2003. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*. 24: 2, 223-242.

O'Donnell, B. Matthew. Römer, Ute. Ellis C. Nick. 2013. The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics.* 18: 1, 83-108

Paquot, Magali, Sylviane Granger. 2012. Formulaic Language in Learner Corpora. 32, 130-149.

Rahusaar, Anne. 2019. The compilation of the spoken sub-corpus for the Tartu corpus of Estonian learner English. Available at http://hdl.handle.net/10062/63917, accessed February 2, 2020.

Simpson, Rita C. et al. 2019. The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan. Available at https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple, accessed May 25, 2020.

Sinclair, John. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.

Tammiste, Lenne. 2016. The use of adjective-noun, verb-noun and phrasal-verb-noun collocations in Estonian learner corpus of English. Available at http://hdl.handle.net/10062/53280, accessed February 15, 2020.

Undo, Aare. 2018. Calculating the Error Percentage of an Automated Part-of-Speech Tagger when Analyzing Estonian Learner English - An Empirical Analysis. Available at http://dspace.ut.ee/bitstream/handle/10062/60466/MA-Thesis_Undo_Aare.pdf?sequence=1&isAllowed=y, accessed May 24, 2020.

Wierzbicka, Anna. 2006. English: Meaning and Culture. Oxford: Oxford University Press.

Wray, Alison. 2000. Formulaic Sequences in Second Language Teaching: Principle and Practice. *Applied Linguistics.* 21: 4, 463-489.

Xiao, Richard and Tony McEnery. 2010. Corpus-Based Contrastive studies of English and Chinese. New York: Routledge.

Appendix 1. Comparison of the spoken corpora

Number	Formula	Freq AFL	Freq TECELE (WC = 21834)	Freq MICASE (WC = 111116)	Freq per 10k TECELE	Freq per 10k MICASE
1	be able to	256	0	8	0	0.7200
2	blah blah blah	29	0	0	0	0
3	this is the	340	0	33	0	2.969869326
4	you know what I mean	64	0	6	0	0.539976241
5	you can see	209	3	27	1.374003847	2.429893085
6	trying to figure out	19	0	4	0	0.3599841607
7	a little bit about	47	0	2	0	0.1799920803
8	does that make sense	29	0	0	0	0
9	you know what	228	0	13	0	1.169948522
10	the university of michigan	35	0	na		0
11	for those of you who	18	0	2	0	0.1799920803
12	do you want me to	14	0	2	0	0.1799920803
13	thank you very much	26	0	0	0	0
14	look at the	197	3	14	1.374003847	1.259944562
15	we're gonna talk about	20	0	0	0	0
16	talk a little bit	19	0	2	0	0.1799920803
17	if you look at	80	0	7	0	0.6299722812
18	and this is	248	0	18	0	1.619928723
19	if you look at the	27	0	3	0	0.2699881205
20	no no no no	31	0	0	0	0
21	at the end of	89	0	6	0	0.539976241
22	we were talking about	23	0	3	0	0.2699881205
23	in ann arbor	19	0	0	0	0
24	it turns out that	24	0	0	0	0
25	you need to	182	0	21	0	1.889916844
26	see what I'm saying	17	0	0	0	0
27	take a look at	31	0	12	0	1.079952482
28	you have a	215	0	17	0	1.529932683
29	might be able to	20	0	0	0	0
30	at the end	137	0	17	0	0
31	you want to	171	0	8	0	0.7199683214
32	to do with	165	0	12	0	1.079952482
33	nothing to do with	22	0	0	0	0
34	know what I mean	65	0	6	0	0.539976241
35	you look at	137	0	13	0	1.169948522
36	university of michigan	44	0	na		
37	what I'm talking about	13	0	0	0	0
38	the same thing	122	0	12	0	1.079952482
39	to look at	131	0	6	0	0.539976241
40	the end of	158	3	13	1.374003847	1.169948522

41	gonna be able to	18		0		0			0		0	
42	we're talking about	61		0		6			0		0.539976241	
43	to figure out what	12		0		1			0		0.08999604017	
44	so if you	170		0		13			0		1.169948522	
45	so this is	173		0		13			0		1.169948522	
46	if you want to	59		0		1			0		0.08999604017	
47	no no no	86		3		2			1.374003847		0.1799920803	
48	if you have	160		0		8			0		0.7199683214	
49	come up with a	17		0		3			0		0.2699881205	
50	we talked about	72		0		7			0		0.6299722812	
51	when you look at	22		0		1			0		0.08999604017	
52	in order to get	23		0		0			0		0	
53	the end of the	88		0		5			0		0.4499802009	
54	oh my god	32		0		0			0		0	
55	come up with	68		0		7			0		0.6299722812	
56	I was gonna say	26		0		2			0		0.1799920803	
57	and then you	170		5		22			2.290006412		1.979912884	
58	a kind of	150		0		33			0		2.969869326	
59	it doesn't matter	51		0		3			0		0.2699881205	
60	has to do with	31		0		2			0		0.1799920803	
61	you can look at	25		0		0			0		0	
62	do you want me	16		0		2			0		0.1799920803	
63	little bit about	48		0		2			0		0.1799920803	
64	if you look	117		0		10			0		0.8999604017	
65	I just wanted to	28		0		3			0		0.2699881205	
66	you're talking about	57		0		12			0		1.079952482	
67	what does that mean	22		0		1			0		0.08999604017	
68	the best way to	18		0		0			0		0	
69	if you want	112		0		7			0		0.6299722812	
70	you know what i	73		0		8			0		0.7199683214	
71	we've talked about	24		0		3			0		0.2699881205	
72	we'll talk about	34		0		1			0		0.08999604017	
73	let me just	44		0		3			0		0.2699881205	
74	I was talking about	14		0		0			0		0	
75	has to be	115		0		9			0		0.8099643616	
76	to talk about	93		3		8			1.374003847		0.7199683214	
77	it turns out	39		0		0			0		0	
78	those of you who	27		0		2			0		0.1799920803	
79	you might want to	19		0		1			0		0.08999604017	
80	first of all	97		0		6			0		0.539976241	
81	and so on and so	17		0		2			0		0.1799920803	

82	there was a	125		5		11			2.290006412		0.9899564419	
83	at the university of	22		0		0			0		0	
84	yes yes yes	30		4		0			1.83200513		0	
85	you can see that	45		0		2			0		0.1799920803	
86	I have a question	31		0		0			0		0	
87	it has to be	37		0		4			0		0.3599841607	
88	we need to	102		0		5			0		0.4499802009	
89	what I'm saying	58		0		1			0		0.08999604017	
90	you want me to	22		0		2			0		0.1799920803	
91	all sorts of	50		0		1			0		0.08999604017	
92	as you can see	20		0		2			0		0.1799920803	
93	to figure out	53		0		6			0		0.539976241	
94	keep in mind	22		0		1			0		0.08999604017	
95	what do you mean	29		0		3			0		0.2699881205	
96	it looks like	66		0		3			0		0.2699881205	
97	let's look at	38		0		5			0		0.4499802009	
98	you look at the	41		0		6			0		0.539976241	
99	to make sure	57		0		4			0		0.3599841607	
100	if you wanted to	19		0		3			0		0.2699881205	
101	make sure that	56		0		3			0		0.2699881205	
102	end up with	38		0		0			0		0	
103	and you can see	39		0		7			0		0.6299722812	
104	came up with	31		0		4			0		0.3599841607	
105	doesn't have to be	17		0		1			0		0.08999604017	
106	I mean if you	41		0		2			0		0.1799920803	
107	you've got a	58		0		3			0		0.2699881205	
108	gonna talk about	41		0		2			0		0.1799920803	
109	how many of you	17		0		0			0		0	
110	I mean if	104		0		7			0		0.6299722812	
111	look at it	80		0		10			0		0.8999604017	
112	piece of paper	16		0		1			0		0.08999604017	
113	and so forth	60		0		2			0		0.1799920803	
114	and you can	142		0		13			0		1.169948522	
115	looking at the	84		0		5			0		0.4499802009	
116	we're gonna talk	23		0		0			0		0	
117	go back to the	22		0		0			0		0	
118	you know what I'm	24		0		1			0		0.08999604017	
119	that you can	136		0		15			0		1.349940603	
120	we're looking at	26		0		4			0		0.3599841607	
121	what I mean	102		0		9			0		0.8099643616	
122	do you know what	31		0		1			0		0.08999604017	

123	how do you know	20		0		0			0		0	
124	you don't need to	20		0		0			0		0	
125	you're looking at	32		0		4			0		0.3599841607	
126	turns out that	28		0		0			0		0	
127	it could be	84		0		21			0		1.889916844	
128	figure out what	26		0		2			0		0.1799920803	
129	if you've got	32		0		0			0		0	
130	I wanted to	84		6		15			2.748007694		1.349940603	
131	you could you could	15		0		1			0		0.08999604017	
132	might be able	20		0		0			0		0	
133	trying to figure	20		0		6			0		0.539976241	
134	what you're saying	40		0		7			0		0.6299722812	
135	we have to	117		0		4			0		0.3599841607	
136	I'm talking about	32		0		1			0		0.08999604017	
137	so you can	114		0		13			0		1.169948522	
138	this kind of	95		4		20			1.83200513		1.799920803	
139	don't worry about	13		0		1			0		0.08999604017	
140	it's gonna be	70		0		6			0		0.539976241	
141	if you have a	45		0		2			0		0.1799920803	
142	wanna talk about	21		0		3			0		0.2699881205	
143	so you can see	18		0		1			0		0.08999604017	
144	I want you to	37		0		3			0		0.2699881205	
145	to look at the	27		0		1			0		0.08999604017	
146	to each other	46		0		4			0		0.3599841607	
147	the kind of	119		0		18			0		1.619928723	
148	at this point	54		0		2			0		0.1799920803	
149	one of these	88		0		4			0		0.3599841607	
150	and if you	132		0		10			0		0.8999604017	
151	you think about it	26		0		4			0		0.3599841607	
152	talk about the	74		0		8			0		0.7199683214	
153	it might be	64		0		7			0		0.6299722812	
154	for those of you	23		0		3			0		0.2699881205	
155	to do with the	43		0		5			0		0.4499802009	
156	I'm not gonna	45		0		5			0		0.4499802009	
157	was talking about	38		0		3			0		0.2699881205	
158	have to do with	20		0		2			0		0.1799920803	
159	tell me what	25		0		4			0		0.3599841607	
160	look at this	57		0		7			0		0.6299722812	
161	in a sense	74		0		10			0		0.8999604017	
162	okay I don't know	14		0		1			0		0.08999604017	
163	I'll talk about	14		0		0			0		0	

164	you need to do	15		0		2			0		0.1799920803	
165	do you want	69		0		3			0		0.2699881205	
166	we talk about	41		0		1			0		0.08999604017	
167	any questions about	14		0		3			0		0.2699881205	
168	come back to	37		0		0			0		0	
169	you can see the	28		0		4			0		0.3599841607	
170	the reason why	36		3		1			1.374003847		0.08999604017	
171	it in terms of	14		0		0			0		0	
172	what I want to	17		0		0			0		0	
173	we looked at	22		0		5			1.374003847		0.08999604017	
174	if you wanna	64		0		10			0		0.8999604017	
175	take a look	41		0		20			0		1.799920803	
176	if you were to	22		0		2			0		0.1799920803	
177	I'll show you	21		0		1			0		0.08999604017	
178	talking about the	64		0		5			0		0.4499802009	
179	that make sense	31		0		0			0		0	
180	this is this is	39		0		4			0		0.3599841607	
181	how do we	59		0		2			0		0.1799920803	
182	we were talking	26		0		3			0		0.2699881205	
183	wanna look at	19		0		0			0		0	
184	you're trying to	38		0		7			0		0.6299722812	
185	a look at	61		0		12			0		1.079952482	
186	if you were to	76		0		2			0		0.1799920803	
187	you're interested in	21		0		3			0		0.2699881205	
188	to think about	81		0		9			0		0.8099643616	
189	gonna be able	18		0		0			0		0	
190	by the way	65		0		5			0		0.4499802009	
191	we look at	43		0		4			0		0.3599841607	
192	I think this is	26		0		1			0		0.08999604017	
193	but if you	94		0		6			0		0.539976241	
194	at some point	24		0		2			0		0.1799920803	
195	I'm gonna go	24		0		3			0		0.2699881205	
196	thank you very	27		0		0			0		0	
197	can look at	34		0		3			0		0.2699881205	
198	what happens is	40		0		2			0		0.1799920803	
199	on the board	30		0		0			0		0	
200	um let me	17		0		2			0		0.1799920803	

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Andreas Piiri

A CORPUS BASED STUDY OF FORMULAIC LANGUAGE USE BY NATIVE AND NON-NATIVE SPEAKERS

Korpuse põhine uurimus formulaarse keelekasutuse kohta emakeelsete ja mitte emakeelsete kõnelejate seas

bakalaureusetöö
2020
Lehekülgede arv:

Annotatsioon:

Märksõnad: Inglise keel ja keeleteadus, formulaic language, corpus analysis, learner language

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavakstegemiseks

Mina, Andreas Piiri,

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

A CORPUS BASED STUDY OF FORMULAIC LANGUAGE USE BY NATIVE AND NON-NATIVE SPEAKERS,

mille juhendaja on Jane Klavan,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpacekuni 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.

Andreas Piiri 26.05.2020

Autorsuse kinnitus

Kinnitan, et olen koostanud käesoleva bakalaureusetöö ise ning toonud korrektselt välja teiste autorite panuse. Töö on koostatud lähtudes Tartu Ülikooli maailma keelte ja kultuuride kolledži anglistika osakonna bakalaureusetöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

[Autori allkiri]

Andreas Piiri 26.05.2020

Lõputöö on lubatud kaitsmisele.

[Juhendaja allkiri] Jane Klavan