UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH STUDIES

Similarities and Differences in the Syntactic Complexity of Bachelor's and Master's Thesis: A Comparative Study Using L2SCA

BA thesis

Kadi Haamer
SUPERVISOR: ***Assoc. Prof.*** **Jane Klavan**

TARTU
2022

ABSTRACT

This thesis investigates the changes in the syntactic complexity of students' written use of language through their years in the university. The basis for the comparison are BA and MA theses from the Department of English Studies in the University of Tartu, Estonia. The research focuses on syntactic complexity measures, which are believed to be good indicators of EFL students' language use. Based on previous research, it is assumed that higher results would indicate a better use of written language.

In the introduction, the importance of theses for the students and their higher level in English are discussed. Additionally, some important terms such as computational tools and learner language are explained. The first section covers previous literature in the field. These topics include: tools that can calculate syntactic complexity, possible shortcomings of these measures, studies that have compared university level EFL student development, and studies that have compared university level students' writing to native English writing. The second section covers the empirical parts of the thesis. First, a Python code is created to clean the MA theses to correspond with learner language standards. The MA theses are then added to an existing corpus of similarly processed BA theses. Next, L2SCA is used to calculate six measures of syntactic complexity for both the BA and MA theses. In addition, the mean number of words and sentences are inspected to identify differences inside BA and MA categories. The gathered results are explained and compared to previous research, which highlights differences between Estonian and other non-native backgrounds, including native English. The section ends with recommendations for future research topics.

Keywords: L2SCA, syntactic complexity, learner language, learner corpus research, bachelor's theses, master's theses, comparative study, Department of English Studies, University of Tartu

TABLE OF CONTENTS

INTRODUCTION

A thesis written at the end of a period of studies such as the bachelor's (BA) and master's (MA) of a certain field can be considered the culmination of what the person has learned during the years. For an English major, this culmination is not only in the knowledge they have acquired in the field of English language and literature but also indicative of their development in the written form of the language. Unlike many students who have only started learning a foreign language when they started their program in the university, English as a foreign language (EFL) learners in Estonia have a considerable knowledge of the language before they start their studies in the university. For example, in 2020 the acceptance into the English Language and Literature BA program of the University of Tartu included the following criteria:

1) English language state examination – min. score 95 (exam taken in 2016–2020) [equal to levels B1 – B2 from CEFR (SA Innove 2018)];
2) Cambridge English: C1 Advanced – min. 200 points;
3) Cambridge English: C2 Proficiency – all levels;
4) The International English Language Testing System (IELTS) Academic – minimum average score 7.5, minimum score for each part of the test separately 7 points;
5) Online Test of English as a Foreign Language (TOEFL) – minimum score 100 (Decision made by TÜ Senate 2020)

This means that gauging the level of the students' knowledge of written English is not as simple as finding mistakes and critiquing their word selection. Instead, to understand a person's development in a language at an advanced level, it is important to identify patterns and writing styles that characterize the student's writing. For this, I have looked comparatively at the use of syntactic complexity elements in BA and MA theses of English majors. My assumption is that during the two years students spend in the MA program of the Department of English Studies their use of written language develops.

Most large quantitative studies of this century have used computational tools in their evaluation of written English (e.g. Lu and Ai 2015, Crossley and McNamara 2012). A computational tool in linguistics is usually a computer program or a system of many

programs that work together to calculate values that can later be used to analyse the text. For example, a computational tool can identify certain parts of the text, such as word classes and sentence structures, and then do calculations based on the values it received. This not only makes the process of evaluating a large number of texts much faster, but also helps in identifying repeating patterns and common mistakes that might otherwise be overlooked.

A specific field within EFL studies is 'learner language', a particularly prominent field in learner corpus research (Granger 1994, Gilquin and Granger 2015, Gilquin et al. 2007). This field takes a quantitative approach and applies computational tools in analysing the language used by learners. What makes learner language data different from other forms of writing in English is that "the focus in learner corpus data is on message conveyance and the possibility for learners to use their own wording" (Gilquin and Granger 2015: 1). It means that learner language research centres around recording the students' own word selection and language use. However, when a computer program goes over a text and calculates the length of sentences and which types of words are used, it will read every part of the file as an element of the student's writing. Having a text which only has learner language therefore means that the document must previously be 'cleaned' from all other data such as page numbers, tables, quotes, and citations that are not part of the students' natural expression in the language. Similarly, the focus of this thesis is not on finding mistakes or evaluating the language used by the students. Instead, it addresses patterns on a syntactic level that might reveal something about the Estonian EFL students' language use.

The aim of this thesis is to identify which differences can be observed in the syntactic complexity of BA and MA theses of the Department of English studies of the University of Tartu. As mentioned before, a thesis is the expression of cumulative knowledge that a student has acquired during their years of study. Finding differences in the sentence structure of their theses could give us an understanding of how much the students' language use has changed

during their studies. There is an expectation that at least in some categories significant differences can be observed between BA and MA theses. It is important to mention, however, that this study does not follow the development of individual students but rather looks at the overall results for both BA and MA theses.

1 Calculating Syntactic Complexity in University-Level EFL Students' Writing

Today, computational tools are a necessity for researchers to conduct large scale research into student writing and language development. For example, there are tools developed specifically for language studies which can assess syntactic complexity and other such measures (Lu 2010, Graesser et al. 2004). Due to the scope of this thesis, existing tools need to be used in the comparison of student texts. Section 1.1 investigates which computational tools are available for assessing language development in non-native student writing. Section 1.2 looks at previous quantitative studies that have similarly described university-level students' development in written language. Section 1.3 explains the process and criteria of choosing syntactic complexity indices that work best for this research.

1.1 Computational Tools and Syntactic Complexity

The availability of computational tools in language studies has changed the way language development in the field of English as a foreign language (EFL) is studied. Previously, most studies in language development were based on a few dozen student texts at a time and only a limited number of measures were used in the analysis (Ortega 2003). On the contrary, most quantitative language studies within the field of language development in the last decades have relied heavily on computer scripts and computational tools (e.g. Ai

and Lu 2013, Crossley and McNamara 2014). This makes it feasible to go through hundreds of student texts at one time. There might still be limitations in these studies, such as the studies only revolving around students from one school or one language background, but the number of texts and measures that can be analysed has grown exponentially.

According to Ortega (2003), one of the best measures for assessing language proficiency in non-native writing is syntactic complexity. She states that syntactic complexity measures are an indicator of the person's ability to use more sophisticated grammar and language (Ortega 2015: 82). Lu and Ai (2015: 17) have defined syntactic complexity as "the range of syntactic structures that are produced and the degree of sophistication of those structures". Byrnes et al. (2010) had a more detailed explanation. While defining syntactic complexity as "the scope of language processing that can be captured within discrete syntactic structures such as sentences and clauses" (Byrnes et al. 2010: 161), they go on to explain that it is natural in all language advancement, whether child or adult, to weave together exceedingly difficult words, phrases, and language patterns based on their obtained knowledge. The ability of the person to recognize and use different sentence structures depends on their understanding of the syntactic elements in the target language (Byrnes et al. 2010: 161).

There are two computational tools that are available for researchers for calculating syntactic complexity – L2 Syntactic Complexity Analyzer (Lu 2010) and Coh-Metrix 3.0 (Graesser et al. 2004). While creating the L2 Syntactic Complexity Analyzer (L2SCA), Lu (2010: 478) used a list of syntactic complexity measures Ortega (2003) had previously studied in detail, a selection of indices from Wolfe-Quintero et al. (1998) and a few from other studies. The L2SCA produces 14 indices of syntactic complexity in categories such as 'length of production unit', 'amount of subordination', 'amount of coordination' and 'degree of phrasal sophistication' (Ai and Lu 2013). Coh-Metrix (McNamara n. d.) has seven indices

of syntactic complexity which can be grouped under the number of words and modifiers, minimum editorial distance scores and syntactic structure scores. It is also possible to find the overall number of specific patterns and the types of words and phrases in the text. However, the focus of Coh-Metrix is not on syntactic complexity. Instead, Coh-Metrix aims to investigate 'cohesion' which is described as features of written language which aid the reader associate their understanding with the text (McNamara n. d.). This means that the tool also provides many other indices such as 'text easability', 'referential cohesion', 'lexical diversity', 'readability' and so forth (McNamara n. d.).

Not all authors agree with the list of syntactic measures proposed by Ortega (2003) and Wolfe-Quintero et al. (1998). There is evidence that some syntactic complexity measures, specifically T-units – a sentence structure consisting of one independent clause and any number of dependent clauses connected to it – could be considered representative of conversational aspects of English (Biber et al. 2011). Similarly, Crossley and McNamara (2014) discovered that human judgements on proficiency did not necessarily match the ratings of the program. In their study human raters preferred clausal features, often indicative of conversational language, to the generally accepted measures of good writing such as nominal style and phrasal complexity. However, the authors cautioned that these results could be indicative of the type of task, the small sample size, and a relatively brief timespan (Crossley and McNamara 2014). These studies indicate a need to be careful in drawing conclusions between high syntactic complexity scores and writing proficiency.

1.2 Previous Studies Using Computational Tools to Measure Syntactic Complexity

Syntactic complexity is often studied in connection to EFL students' improvement over time. One very recent example is a longitudinal study by Polat et al. (2020) based on

Turkish learners of English. They studied a corpus of university level essays and followed the progression of students over three proficiency levels – elementary, pre-intermediate and intermediate. Using the abovementioned tool L2SCA, the authors calculated the syntactic complexity of student writings and found that instead of the linear improvement that is often expected, their results were not as predictable. Ten out of the 14 indices indeed had a linear growth through the levels, though even then the significance could be seen between the first and the third levels but not when comparing them level by level (Polat et al. 2020). This mirrors the findings of Lu (2011) who analysed the syntactic complexity of Chinese EFL students between four university levels using the same L2SCA tool. He explained that even though in many instances there were no significant differences in adjacent years, development could be seen over a longer period. This was also confirmed by Ai and Lu (2013) who combined Chinese students' essays from years 1-2 and 3-4 and could see a larger difference between the lower and higher levels than was visible in Lu's previous study (2011).

However, not all indices showed a linear increase. Polat et al. (2020) discovered that the remaining four indices (T-unit complexity ratio, verb phrases per T-unit, coordinate phrases per T-unit and Coordinate phrases per clause) either had no difference between the first two levels and then had a higher result in the third, an initial increase between first and second level and later a decrease in third, or the results decreased over time. This once again is reflected in the results obtained by Lu (2011), who noticed three instances where the change was negative when examined over a longer period – sentence complexity ratio, dependent clause ratio, and dependent clauses per T-unit. It is interesting to note that none of these 'negative' categories were the same across the two studies. These results show that progress in a language is not always fast, predictable, or linear in nature (Polat et al. 2020).

Much fewer articles have been published in recent years that compare the syntactic

complexity of non-native speakers' (NNS) writing to native English speakers' (NS) writing, especially at university level. There are two studies that stand out: Ai and Lu (2013) and Azadnia et al. (2019). Both studies used one of the above-mentioned tools to compare syntactic complexity measures in NNS and NS texts at university levels. Using the L2SCA tool, Ai and Lu (2013) concentrated on the syntactic differences between Chinese student essays of different university levels and compared them to NS essays from the Louvain Corpus of Native English Essays (LOCNESS; Granger 1998). They found that in all categories of syntactic complexity, the NNS essays returned lower results than the NS essays (Ai and Lu 2013). The only two categories showing no significant difference were related to coordination – coordinate phrases per clause and T-units per sentence. Azednia et al. (2019) used the Coh-Metrics tool for their analysis but had similar results when comparing Iranian Ph.D. theses and native English dissertations. One noticeable difference was the high number of modifiers in the Iranian EFL texts which the authors believe could be explained by Ph.D. level students having more experience with academic writing or the foreign authors putting more effort into trying to be understood than their native counterparts.

An important topic to understand possible differences between previous studies and Estonian EFL learner results is the background language of students. Lu and Ai (2015) performed an interesting study where they sampled 200 argumentative essays from seven different first language (L1) backgrounds (ICLE 2.0; Granger et al. 2009) and compared them to 200 essays of native English origin (LOCNESS; Granger 1998). Using the tool L2SCA, they looked at the syntactic complexity of Japanese, Chinese, Bulgarian, Russian, Tswana, French, and German college-level student essays. When looking at the data altogether, only three categories had identifiable differences – mean length of clause, complex nominals per clause and complex nominals per T-unit. However, when separately comparing the NNS essays of different backgrounds to NS essays, the study found that there

were some significant differences in all 14 categories (Lu and Ai 2015). This is contrary to previous findings where certain categories of syntactic complexity were deemed to be irrelevant based on them not having any significant differences between students' levels or in comparison to NS essays (Lu 2010). It is then necessary to consider the first language of EFL learners. Looking at the writing habits in their native language can be an indication of why certain syntactic patterns can be observed in their English writing.

1.3 Selecting the Measures of Syntactic Complexity

The computational tool that was selected for the processing of the theses was the L2 Syntactic Complexity Analyser by Lu (2010). The focus of this analysis is on measures of syntactic complexity. The L2SCA can calculate more types of syntactic complexity measures and was therefore chosen over Coh-Metrix (Graesser et al. 2004). There was also a broader selection of studies that used the L2SCA indices for measuring syntactic complexity, making it possible to learn from their results and methods (Ai and Lu 2013, Polat et al. 2020, Lu and Ai 2015).

As mentioned before, L2 Syntactic Complexity Analyzer (Lu 2010) was created specifically to identify syntactic elements in written text. It can calculate 14 indices of syntactic complexity:

1. Mean length of clause
2. Mean length of sentence
3. Mean length of T-unit
4. Sentence complexity ratio
5. T-unit complexity ratio
6. Complex T-unit ratio
7. Dependent clause ratio
8. Dependent clauses per T-unit
9. Coordinate phrases per clause

10. Coordinate phrases per T-unit
11. Sentence coordination ratio
12. Complex nominals per clause
13. Complex nominals per T-unit
14. Verb phrases per T-unit (Lu 2010)

These measures can be grouped roughly into four categories: mean length of components (1-3), subordination between components (5-8), coordination between components (9-11), and other constructs (4, 12-14). In addition to these results, the program also provides the number of words, sentences, clauses, dependent clauses, T-units, complex T-units, coordinate phrases, complex nominals, and verb phrases it counted in the process (Lu 2010).

To fit in the constraints of this thesis the number of measures that are analysed were limited. Due to previous research identifying T-units as more indicative of spoken English rather than written (Biber et al. 2011), all measures that include T-units were discarded. Ultimately six indices of syntactic complexity were included in the analysis – mean length of clauses, mean length of sentences, ratio of dependent clauses, coordinate phrases per clause, complex nominals per clause, and sentence complexity ratio. The explanation for these measures can be found in section 2.2, which covers the process of calculating syntactic complexity measures using L2SCA.

2 Analysis of Bachelor's and Master's Theses from the Department of English Studies Using Syntactic Complexity Measures

The empirical section of this study covers the method used and the analysis of the results gained in the practical part of this thesis. To compare the BA and MA theses, the texts had to first be converted into TXT format and cleaned to only contain learner language. A

student from the previous year (Kaljuste 2021) already converted the BA theses into the proper format and created the corpus Estonian Academic Learner's English (EALE). Although the final method used for cleaning the data in the present study differed from the method used by Kaljuste (2021), the same principles were followed in converting the MA theses. After the files were correctly converted and cleaned, they were run through the L2 Syntactic Complexity Analyzer (Lu 2010). The program counted the frequency of certain structures in the text and then calculated 14 syntactic complexity indices. Out of these, six measures were chosen for this analysis. Once the results were calculated from both bachelor's and master's theses, they were analysed and compared to each other.

Section 2.1 covers the process of formatting and cleaning the MA thesis of the Department of English studies. Sections 2.2 and 2.3 explain the results gathered from L2SCA (Lu 2010). Section 2.4 compares the results between bachelor's and master's theses and examines the findings in comparison to other similar studies. Section 2.5 discusses the findings and proposes directions for future research.

2.1 Formatting and Cleaning the Theses

The MA theses of the English language and literature speciality and English teacher speciality are uploaded annually to the University of Tartu's DSpace in PDF format. As of this year there were 162 submissions in the MA thesis database. However, the aim of this research was to compare the BA and MA theses, and the practice of writing BA theses began in 2018 in the Department of English Studies. Therefore, only MA theses from 2018 to 2020 were selected for this research, amounting to 58 texts. Theses from the year 2021 were not included in the selection since Kaljuste's (2021) corpus only had BA theses until the year 2020.

Each thesis was downloaded and converted from PDF format into TXT format by

using an online tool – Easy PDF (2022). There were two exceptions that the chosen program was unable to convert. For these PDF2Go (2022) was used instead. In this step I deviated from Kaljuste's (2021) method, who used a code in the programming language R (version 4.0.3, R Core Team 2020) for both formatting and cleaning the texts. In the description of his method Kaljuste (2021: 17) mentioned that a lot of junk data was added in the process of converting the files from PDF to TXT when using R, which in turn required further steps to remove the additional noisy data. Since the PDF files in my selection were relatively new and machine-readable, only a simple file conversion was needed. There are many existing online tools that can do it, and in this instance Easy PDF was selected because it did not have a limit on the number of conversions that could be made using the free version.

The next part of the process was cleaning the texts so that only the learner language portion of it remained. Kaljuste (2021) performed the cleaning of BA theses in the programming language R. Since there were some shortcomings in his code that required manual fixing later, it was decided to create a code in the programming language Python 3.9 (Python Software Foundation 2020) for the purposes of the present study. Following the example of Kaljuste (2021), the code first identified and removed everything except the abstract and the main body of the text. Next, the code went over the document and removed page numbers and any tables and figures contained in the document. Finally, the code removed all quotations and citations in the text as well as any empty space these changes created. This process removed any parts from the texts that would otherwise obstruct analysing the language used by the student.

Something that was done differently from Kaljuste (2021) was that each of the sections that would be removed were first highlighted. In case a section was wrongly identified by the program it could either be marked as an exception or corrected in the original file before continuing. This was especially important for tables since the program

often identified half of a table while leaving some rows untouched, or erroneously identified other sections as a table due to additional space between words. Fixing these sections manually took time, but also helped to make sure that any anomalies created during the conversion process were identified and removed from the final text.

It was decided that any information longer than two words in between quotation marks would be considered a quote. This ensured that if a person had highlighted a term or word by using quotation marks it would not be automatically removed from the text. For brackets the criterion was that anything that included a number was considered a citation. Even though there was a possibility to manually add exceptions, the instances where people added examples or additional information into brackets that did not follow the criteria were too numerous to add each instance into the list of exceptions. Thus, only repeated phrases or important terms (such as L1 and L2) were counted among the exceptions.

During the process of going over the theses some imperfections of the Python code were identified. The most frequent problem occurred when the writer accidentally used the same quotation marks in the end of the quote as in the beginning (“/“). It was even more problematic when the author mixed different types of quotation marks in their thesis, for example, using Estonian quotation marks („/“) and English quotation marks (“/”) interchangeably. In both instances the program identified the end of one quote as the beginning of another. It was also problematic when a person used the same brackets or quotation marks inside and outside of the phrase. The code was built so that after it encounters the beginning of a bracket or quotation, it finds the first closing symbol of the same kind. This meant that when there were brackets inside brackets, the program only eliminated a portion of the sentence or phrase that was supposed to be removed. These instances all required manually changing the original document before going over the text once more with the program.

Similarly to Kaljuste (2021), there was no solution to removing long citations from the theses as they were formatted identically to regular paragraphs once the files were in TXT format. For this step, each thesis had to be inspected separately and the long citations removed manually. An unexpected dilemma was identifying whether lists were paraphrased by the student, or they were long citations from another author. In many instances, students began the sentence as '[the author] said:' but there were no identifiable markers whether the points were paraphrased or taken directly from the author. In this instance I resolved to remove only sections that were either in smaller font size or had a citation at the end as well. Once the texts had been formatted and cleaned, they were uploaded to the EALE corpus (Kaljuste 2021). The commented code for the cleaning of the MA theses can be found in R. E. Haamer's Bitbucket (Haamer 2022).

2.2 Process of Using L2SCA to Calculate Measures of Syntactic Complexity

The code for the original implementation of the L2SCA, Version 3.3.3 (Lu 2016) was downloaded directly from Lu's website. The code can be used in both MacOS and Linux operating systems and requires both Python and Java programming languages to run. There were also other implementations available which were developed by different authors, such as an online interface, a GUI for Windows and MacOS, and an implementation for the R programming language (Gaillat et al. 2019). For this analysis the original implementation was selected since it could be run from Linux and used the more familiar Python programming language.

The program can go through entire folders at a time, so it was decided to divide the theses into different folders based on years and separate them by whether they were BA or MA theses. At this point the names of the files were changed to protect the authors' identity. Instead, a naming system of BA or MA, followed by the year and an identifying number was

selected, for example MA2019_001. This naming system was also used when the cleaned TXT files were uploaded to the EALE corpus (Kaljuste 2021). Since the BA theses in Kaljuste's corpus did not include years, the file names were changed manually to reflect the naming system. Even though the theses were divided into different years, for the purpose of this study all years were viewed as one corpus of either BA or MA theses. Having a bigger corpus of each was better for drawing general conclusions about the differences between BA and MA students' language use. The distinction between years was left for future studies that might be interested in comparing student writing through the years or only using specific years in their comparison.

Out of the 14 syntactic complexity indices that the L2SCA tool can calculate, six were selected for this study. As explained in section 1.3, the measures including T-units were left out since they are considered more indicative of spoken rather than written English (Biber et al. 2011). The indices explored in this study are:

1) Mean length of sentence – calculated by diving the number of words by the number of sentences. The program uses the Stanford parser (Klein and Manning 2003) to divide the text into individual sentences, add tokens to each part of the sentence and tag each part-of-speech with a label (Lu 2010: 479-480).
2) Mean length of clause – calculated by diving the number of words by the number of clauses. Lu (2010: 481) defines clauses as "a structure with a subject and a finite verb", this encompasses independent clauses, adjective clauses, adverbial clauses, and nominal clauses. In his code, he uses certain Tregex patterns (Levy & Andrew 2006) to identify all clauses and phrases in the sentence.
3) Sentence complexity ratio – calculated by dividing the number of clauses by the number of sentences.

4) Complex nominal ratio – calculated by dividing the number of complex nominals by the number of clauses. There are three categories that Lu (2010: 483) counts under complex nominals: "(i) nouns plus adjective, possessive, prepositional phrase, relative clause, participle, or appositive, (ii) nominal clauses, and (iii) gerunds and infinitives in subject position".
5) Dependent clause ratio – calculated by diving the number of dependent clauses by the number of clauses. Lu (2010: 482) counts finite adjective, adverbial and nominal clauses under dependent clauses.
6) Coordinate phrases per clause – calculated by dividing the number of coordinate phrases by the number of clauses. Coordinate phrases include adjective, adverb, noun, and verb phrases (Lu 2010: 283).

After receiving the results, averages and standard deviations were calculated for each category to find similarities and differences in the BA and MA theses. In addition to the calculated results above, the program counted the number of words, sentences, verb phrases, clauses, t-units, dependent clauses, complex t-units, coordinate phrases and complex nominals. For each of these measures, values for average, minimum and maximum were calculated to analyse the basic differences between BA and MA theses. A table with all the results can be found in the Appendix. In this study I will concentrate on the six calculated values and the number of words and sentences.

2.3 Explanation of Initial Results

At the Department of English studies of Tartu University there are three different programs under the MA studies. Out of these, two are represented in this corpus – the European Languages and Cultures and the Teacher of Foreign Languages. In addition to the traditionally longer thesis, there is also an option to do a master's project with a written

explanation which are generally shorter than ‘traditional’ theses. This means that there are at least three different criteria for the MA theses in this corpus. It came as no surprise then that the length of MA theses was quite varied.

As can be seen in Table 1, the average word count of the MA theses was 13,600 words and the average sentence count was 580 sentences. However, even knowing the differences in criteria, it was still interesting to see that the difference between the maximum and minimum number of words and sentences was so large. The highest numbers were 26,632 words and 1,197 sentences, while the lowest were 1,933 words and 97 sentences. In

Type of thesis	**Nr. of thesis**	**Total nr. of words**	**Words per thesis**			**Sentences per thesis**		
			AVG	MIN	MAX	AVG	MIN	MAX
MA	57	775,243	13,600.75	1,933	26,632	579.84	97	1197
BA	73	492,285	6,743.63	3,541	11,732	284,86	161	618

Table 1. General information about the theses in the corpus.

both the maximum and the minimum instances, the word and sentence counts came from the same thesis. However, the shortest thesis was an outlier in the corpus with the second shortest thesis producing 7,698 words and another thesis having 315 sentences.

Whereas the master’s theses had different programs represented in the corpus, there is only one program and criteria for theses at the BA level. It was a surprise then to find that these theses also had a very large range in the highest and lowest results. The average numbers for the BA theses were 6,744 words and 285 sentences. The highest number of words per thesis was 11,732 and the lowest 3,541, and the highest and lowest number of sentences were respectively 618 and 152. Unlike the MA theses, the extremes for the BA thesis were not from the same texts. The work with the highest number of words (11,732) had 309 sentences, only slightly higher than the average for BA theses. At the same time,

the work with the highest number of sentences (618) had the second highest word count at 10,574. The thesis with the lowest number of words (3,541) also had quite a low number of sentences at 161, while the work with the lowest number of sentences (152) also had the second lowest number of words at 3,583. Even though the differences between the maximum and minimum numbers were quite large, the BA theses overall had a more even distribution between their numbers than the MA theses.

Table 2 shows the averages (AVG) and standard deviations (SD) of each syntactic complexity measure that was calculated for the BA and MA theses. Looking at the mean lengths of sentences and clauses, the MA theses had an average of 23.83 words in a sentence and 11.66 words in a clause. The standard deviation for the sentences was quite high at 2.99, while the standard deviation for the clauses was a more modest 1.14. For the BA theses, the average length of sentences was 24.23 words (0.4 longer than MA), and for the clauses it was 11.41 words (0.25 shorter than MA). The differences between the BA theses were even more prominent than for the MA theses. The standard deviation for the sentence lengths was 4.76 and the standard deviation for the clause lengths was 1.53.

Syntactic Complexity Measures	**BA**		**MA**	
	AVG	SD	AVG	SD
Mean length of sentence	24.23	4.76	23.83	2.99
Mean length of clause	11.41	1.53	11.66	1.14
Clauses per sentence	2.13	0.32	2.05	0.26
Complex nominals per clause	1.47	0.3	1.51	0.21
Dependent clauses per clause	0.44	0.06	0.42	0.06
Coordinate phrases per clause	0.32	0.09	0.35	0.09

Table 2. Syntactic Complexity Measures in BA and MA theses

The subsequent categories analysed were the clauses per sentence and the complex nominals per clause (Table 2). The MA theses had on average 2.05 clauses per sentence and 1.51 complex nominals per clause. The standard deviations for these calculations were much lower than for the mean length categories – 0.26 for clauses per sentence and 0.21 for complex nominals per clause. The average number of clauses per sentence for the BA theses was 2.13 (0.08 more than MA) and for the complex nominals it was 1.47 per clause (0.04 less than MA). The standard deviations were one decimal higher for the BA theses than for the MA theses – 0.32 for the clauses per sentence and 0.3 for the complex nominals per clause.

The two categories that show the use of subordination and coordination were respectively the dependent clauses per clause and the coordinate phrases per clause (Table 2). For the first, the average use of dependent clauses for the MA theses was 0.42 per clause, while the use of coordinate phrases was 0.35 per clause. The results were also quite consistent based on the standard deviation, both being less than 0.1 – 0.06 and 0.09 respectively. For the BA theses, the average of dependent clauses per clause was very similar to the MA theses with a result of 0.44 (0.02 more than MA). The coordinate phrases per clause for BA theses only had a slightly larger difference with a result of 0.32 per clause (0.03 less than MA). It is noteworthy that the standard deviations for these measures were the same as for the MA theses – 0.06 for dependent clauses and 0.09 for the coordinate phrases.

2.4 Discussion of L2SCA Results in Comparison to Previous Research

In this section, the numbers presented in the previous section (2.2) are looked at in more detail and possible causes for the values are proposed. The calculated values showed both increase and decrease when comparing BA and MA theses. To understand whether this

is a positive or negative effect, some results are compared to numbers from native American English essays that Ai and Lu's used in their comparative study (2013). Even though these essays are written as coursework and are likely not as important to the students as a final thesis, the results can be indicative of the written language used by native speakers of English. The results of the present study are also compared to a study by Lu and Ai (2015) which looks at the influence of L1 on syntactic complexity measures. Additionally, some problems arising from the cleaning process are looked at in more detail to ascertain that the conclusions drawn are not skewed because of faulty data.

One of the interesting aspects of the data above was the length of theses, described by the number of words and sentences. The difference in length was quite surprising for the BA theses since they all came from the same program. The difference between the shortest and longest theses was around 8,200 words, the longer ones being more than three times the length of the shorter texts. While the English studies department of the University of Tartu has had a tradition of writing MA theses for decades, the requirements for the BA theses for the 3-year BA programme were (re-)introduced in 2018. It is possible then that the criteria of what is expected of the students has changed slightly over the years. For example, a clear difference can be seen when looking at the average length of BA theses in 2018 (5,877 words, standard deviation 1,122) and in 2020 (7,227 words, standard deviation 1,653). The theses from 2018 were not only shorter overall, but also differed less in length than the theses from 2020. Small differences in the wording of the guidelines, for example requesting a range (25-30 pages) or an estimate (around 30 pages) of a thesis length could influence the consistency in length. Currently, the requirements of the thesis for each study year are not known to the author.

It was mentioned previously that the MA theses had different criteria and thus different results were expected depending on the specialty the students graduated. These

differences could easily be seen in the word count of the theses – the difference of MA thesis length raged from around 8,000 words (leaving out the thesis with only 1,933 words) to around 25,500 words. There was no way to categorise these MA theses based on specialty since the theses did not have any markers on their title pages, file names or upload locations that could help distinguish the theses from one another. Even though most of the indices calculated were using an average as their base value, it is possible that the theses would have had other distinguishable differences that were overlooked because they were viewed as belonging to one category.

In the process of cleaning the theses, one of the steps was removing quotations from the text so that only the students' own words remained. In some instances, it meant removing the whole sentence together with the citation at the end. In most, however, it meant that a portion of the sentence was left in while the rest was removed. For example, sentences that now consisted of "[the author] said: ." or "While [this author] defined it as , [that author] argued .". It is important to note then that not all the data collected, especially about sentence lengths, is completely indicative of the author's natural writing. However, since the study focuses on learner data and learner language, removing quotations was vital from the viewpoint of the study.

This brings us to the next category, mean lengths of sentence and clause. Mean length of sentence was one of the values that surprisingly decreased from BA and MA theses (24.23 and 23.83 respectively). This could be indicative of MA students using more quotations that are integrated into their own sentences. After cleaning the theses, there would be many sentences that consisted of only a few words, lowering the overall length of sentences. The decreasing number at higher levels could also be indicative of students being clearer and more precise in their expression. This assumption is also supported by the standard deviations – for the BA theses, the standard deviation was 4.76 and for the MA theses, it was

2.99. This shows more consistency in the lengths of MA theses' sentences.

However, looking at Ai and Lu's (2013) university level native speaker (NS) results, their mean sentence length was 19 words. This much smaller number for NS raises the possibility that the Estonian language background could influence student's writing in English. In a study where non-native (NNS) first language backgrounds were observed, about half of the languages (Japanese, Chinese and Russian) had a lower average sentence length than 19 and the rest (Bulgarian, French, Tswana and German) had slightly higher results (Lu and Ai 2015). However, even German essay values, which on average had the longest sentences with a mean of 22.31, were lower than the Estonian values at 24.23 and 23.83. This data for other language backgrounds came from college level argumentative essays, which is not fully comparable to a thesis, but could give a rough estimate of which results other NNS backgrounds are likely to produce. If students from all language backgrounds tend to use shorter sentences than Estonians, it is very likely that the Estonian language background and/or Estonian academic writing conventions are influencing this tendency. An important aspect to bear in mind, at this point, is that there are some foreign students each year in the European Languages and Cultures master's program (around 2 per year); this could have influenced the on average shorter sentences in the MA theses and other results.

The mean length of clause was similar for both, showing only a 0.25 increase for MA theses and a slightly smaller standard deviation than for BA theses. However, when comparing the BA and MA theses (11.41 and 11.66) to NS writing, the average length of clauses for Estonians was about 1.5 words longer than in the NS essays (9.94) (Ai and Lu 2013). This is an indication that the increase from BA to MA level might have in this instance been unfavorable. At the same time, the number of clauses per sentence, although by only a small amount (0.08), was once more less for MA theses than for BA theses. One possibility

for this is that compared to native speakers, EFL students are more likely to add too much information to their sentences. An example of that is a run-on sentence – a sentence with two or more independent clauses linked without any or erroneous punctuation marks and conjunctions (Zheng and Park 2013). By having a smaller number of clauses, the intended meaning of the sentence could become clearer. This is supported by the evidence gathered from NS essays, where on average they used less than two clauses per sentence (1.97) (Lu and Ai 2015).

There were three values calculated per clause – complex nominals, dependent clauses, and coordinate phrases. All of these showed very little change between BA and MA theses, less than 0.05 for each. For all the other categories, MA theses showed a decrease in the standard deviations, indicating a more consistent result between all the students. The standard deviations, however, were almost the same between MA and BA theses for all these categories, only complex nominals showed 0.09 less deviation for MA theses than for BA.

Complex nominals and coordinate phrases per clause were two of the categories that showed a very slight increase for MA students. However, NS and other NNS essays had again much lower numbers. For complex nominals, the results started from 0.91 for Tswana and were the highest for French students with 1.28 (Lu and Ai 2015). Even though the NS English results were on the higher side (1.25) (Ai and Lu 2013), the Estonian students' results (1.47 for BA and 1.51 for MA) were still much higher in comparison. The same can be said for coordinate phrases per clause – the NS result was 0.25 (Ai and Lu 2013) and the rest of the NNS ranged from 0.17 (Japanese) to 0.3 (Bulgarian). The result for Turkish learners was even lower at 0.12 (Polat et al. 2020). Compared to those, the results of 0.32 for BA theses and 0.35 for MA theses were quite a bit higher than the average for other language backgrounds.

The category that showed the slightest difference between BA and MA theses was

the dependent clauses per clause category. For BA theses the mean was 0.44 and for MA theses it was 0.42. Compared to other backgrounds (lowest 0.35 for Chinese and highest 0.46 for Tswana), these results were quite average and close to the NS result of 0.4 (Lu and Ai 2015).

2.5 Discussion of Results

In their study Ai and Lu (2013) looked at lower and higher levels of university level Chinese students' writing. When they compared their results to native English essays, they saw almost a linear growth between the NNS levels and NS results. In the comparative study by Polat et al. (2020) similar tendencies could be seen from their results, even if they were not as consistent as Ai and Lu's (2013) results. In the study by Lu and Ai (2015) when they compared all the NNS essays as a group to the NS essays, the results were mostly lower for the NNS group. Following these studies, the assumption before starting this comparison was that in most categories, higher results would equal higher proficiency in the language. In contrast, it could be said that for Estonian university level students, lower results in syntactic complexity measures equals better language use.

Estonian students in both BA and MA levels tend to write longer and more complex sentences than their European and native English counterparts. There were three European languages that were included in the study by Lu and Ai (2015) – Bulgarian, French and German. Out of these, Bulgarian and French essays were usually quite close to the native English texts, and German essays tended to have higher results across all measures. However, even in comparison to German results, both BA and MA theses of Estonian students had remarkably high numbers. The only categories that were not significantly higher than the results of German EFL learners were clauses per sentence and dependent clauses per clause. Based on this study, syntactic complexity is a field that should be more

closely looked at by EFL teachers in Estonia – do Estonians tend to overcomplicate their sentences or use needlessly long sentences at higher levels of proficiency? This remains an important research question for further studies in this field. Hopefully, in the future, a corpus of Estonian university level writing can be created, where the data consists of course essays rather than theses. This would provide a better ground of comparison to existing native English student writing.

Finally, it is important to note that not all these results are the full reflection of Estonian students' writing. Theses are written over a long period of time and not in a controlled setting. Students might get help from their friends who have a higher proficiency in the language, or in the worst-case scenario someone else could write the thesis for them. Even if they do not get excessive help from outside, they likely get comments from their supervisor in how to improve their text and perhaps even how to correct their sentences. When comparing the BA and MA theses to each other, the proportion of assistance from supervisors is likely similar in both levels. However, this needs to be considered when comparing these theses to other types of writing. There is also the added factor that not all students in the BA or MA level have Estonian as their first language. As discussed in the previous section, students' L1 background can have a considerable influence on the way they form their sentences in English.

In the future, it would be beneficial to compare Estonian EFL theses to native English theses of similar levels. As mentioned before, the resources for both NS and NNS backgrounds were college level essays (Ai and Lu 2013, Lu and Ai 2015). It is possible that the higher overall results for Estonian theses were not as much a question of language background but the type of writing task. The theses that were looked at in this study also had students of different language backgrounds other than Estonian which might have influenced the overall results. Eliminating the difference in task type and focusing on only students from

one language background could better show which language patterns are common for high level Estonian EFL learners in comparison to native English writing.

Another interesting research topic could be analysing the individual development of students between BA and MA levels. Since the first BA theses were written in 2018, there were only a couple of students that had finished both their BA and MA studies in this department by 2020. However, each year there will be more students graduating who have written both a BA and MA thesis in this department. It could be very interesting to see how much students' writing has actually changed during the two years of study.

CONCLUSION

Syntactic complexity is often used in EFL studies to assess students' development in the language. There are studies that caution against equating certain syntactic complexity results to proficiency levels, and doubts have been raised about whether some accepted sentence structures are more indicative of conversational or written language (Crossley and McNamara 2014, Biber et al. 2011). Even if higher numbers in these measures might not be indicative of proficiency levels, syntactic complexity measures are still invaluable in comparing possible tendencies in language use.

This study looked for similarities and differences between BA and MA theses of the Department of English Studies of the University of Tartu using measures of syntactic complexity. These theses were converted into the proper format, cleaned to follow conventions of learner language collection (Kaljuste 2021, Haamer 2022) and put through the L2SCA tool (Lu 2010). Values such as the number of words, sentences and clauses were collected for each thesis. In addition, six values were recorded from the 14 that the L2SCA could calculate. These six measures include the following: mean length of sentence, mean length of clause, clauses per sentence, complex nominals per clause, dependent clauses per clause and coordinate phrases per clause. For each measure, averages and standard deviations or minimum and maximum values were calculated to better understand the differences between BA and MA theses' results.

It was found that both BA and MA theses had a wide range of lengths where the longer texts had about three times as many words as the shorter texts. While this was understandable for the MA theses since they came from different specialities with different requirements, the results were unexpected for the BA theses. It was suggested that the difference in BA theses' lengths could come from the relatively new tradition of writing BA theses in the English Language and Literature program and possible small differences in the

expectations based on which year the students have written their theses.

The next result that was looked at was the mean length of sentences. This showed a slightly surprising difference, where the sentences used by BA students tended to be longer than for MA students. Three possible explanations were proposed. First, that the MA students used more in-text quotations that were removed during the cleaning process and thus lowered the overall average length of sentences. Second, that they were more precise and consistent in their expression. This was supported by the MA theses having a smaller standard deviation than BA theses. Finally, the sentence lengths were compared to native English essays as well as essays from other backgrounds. All of them had a smaller mean sentence length than Estonian students, showing that the lower result at higher levels might have been a change in the right direction, if we take native English writing as the standard to draw comparisons with.

The next categories that were looked at were mean length of clause and clauses per sentences. In the mean length of clause, the MA theses had a slightly higher result than for BA theses. For the clauses per sentence measure, the results were very similar for both MA and BA theses, showing only a slight decline in the MA level. However, for native speakers, the average clause length was about 1.5 words less and they tended to use less clauses than Estonian students. A possible explanation to this could have been that EFL students tend to add too much information to their sentences, which would also add clauses. Lastly, three categories were looked at that showed almost no change between BA and MA theses – complex nominals, dependent clauses, and coordinate phrases. However, other than complex nominals per clause, they were still higher than the results for both native essays and essays by other language backgrounds.

This study showed that overall, Estonian theses had higher results in almost all calculated syntactic complexity categories compared to essays from other language

backgrounds. Unlike theories by Ai and Lu (2013) and Polat et al. (2020), who assumed that linear improvement in all categories show a better use of written language, for Estonians this could be the opposite. Further research into this field is required to eliminate the possibility of the type of task or mixed language backgrounds influencing the results.

REFERENCES

Ai, Haiyang and Xiaofei Lu. 2013. A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. *Automatic treatment and analysis of learner corpus data*, 249-264.

Azadnia, Masoud, Ahmadreza Lofti and Reza Biria. 2019. A Study of Syntactic Complexity via Coh-Metrix: Similarities and Differences of Ph. D. Dissertations Written by Iranian University Students and English Native Speakers. *Research in English Language Pedagogy*, 7: 2, 232-254.

Biber, Douglas, Bethany Gray and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *Tesol Quarterly*, 45: 1, 5-35.

Byrnes, Heidi, Hiram H. Maxim and John M. Norris. 2010. Realizing Advanced Foreign Language Writing Development in Collegiate Education: Curricular Design, Pedagogy, Assessment. *The Modern Language Journal*, 94: i–235. Available at: http://www.jstor.org/stable/40985261, accessed April 16, 2022.

Crossley, Scott A., Danielle S. McNamara. 2012. Detecting the First Language of Second Language Writers Using Automated Indices of Cohesion, Lexical Sophistication, Syntactic Complexity and Conceptual Knowledge. *Approaching language transfer through text classification: Explorations in the detection-based approach*. Channel View Publications: 106-126.

Crossley, Scott A., Danielle S. McNamara. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66-79.

Easy PDF. 2022. *PDF to Text*. Available at http://www.easypdf.com/pdf-to-text, accessed February 22, 2022.

Gaillat, Thomas, Nicolas Ballier. 2019. Expérimentation de Feedback Visuel Des Productions écrites d'apprenants Francophones de l'anglais Sous MOODLE [Visual Feedback Experiment in Written Productions of French-Speaking Learners of English In MOODLE]. *Actes de La Conférence EIAH2019*. Paris: Association des Technologies de l'Information pour l'Education et la Formation.

Gilquin, Gaëtanelle, Sylviane Granger, and Paquot, M. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6: 4, 319-335.

Gilquin, Gaëtanelle, Sylviane Granger. 2015. Learner Language. In Douglas Biber and Randi Reppen (eds). *Cambridge Handbook of Corpus Linguistic*, 418-435. Cambridge: Cambridge University Press.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36: 2, 193-202.

Granger, Sylviane. 1994. The learner corpus: A revolution in applied linguistics. *English Today*, 39: 3, 25-29.

Granger, Sylviane. 1998. The computer learner corpus: A versatile new source of data for SLA research. In Sylviane Granger (ed). *Learner English on Computer*, 3-18. London and New York: Addison Wesley Longman.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot (eds.). 2009. *International corpus of learner English*. Vol. 2. Louvain-la-Neuve: Presses universitaires de Louvain.

Haamer, Rain Eric. 2022. Available at: https://bitbucket.org/erichaamer/kaditextit66tlus/src/master/, accessed on May 21, 2022.

Kaljuste, Karl August. 2021. *Error Rate of Automated Part-of-speech Tagging of Estonian Academic Learner English*. Unpublished BA thesis. Department of English Studies, University of Tartu, Tartu, Estonia.

Klein, Dan, Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun & K. Obermayer (eds). *Advances in Neural Information Processing Systems*, 15: 3–10. Cambridge: MIT Press.

Levy, Roger, Galen Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. *LREC*, 2231–2234.

Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15: 4, 474–496.

Lu, Xiaofei. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL quarterly*, 45: 1, 36-62.

Lu, Xiaofei. 2016. *L2 Syntactic Complexity Analyser Version 3.3.3*. Available at http://www.personal.psu.edu/xxl13/downloads/l2sca.html, accessed 13.05.2022.

Lu, Xiaofei, Haiyang Ai. 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29, 16-27.

McNamara, Dr. Danielle. N. d. Coh-Metrix version 3.0 indices. Available at: http://cohmetrix.memphis.edu/cohmetrixhome/documentation_indices.html, accessed December 23, 2021.

Ortega, Lourdes. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, 24: 4, 492-518.

Ortega, Lourdes. 2015. Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, 29, 82-94.

PDF2Go. 2022. *Convert PDF To Text*. Available at https://www.pdf2go.com/pdf-to-text, accessed March 23, 2022.

Polat, Nihat, Laura Mahalingappa, Rae L. Mancilla. 2020. Longitudinal growth trajectories of written syntactic complexity: The case of Turkish learners in an intensive English program. *Applied Linguistics*, 41: 5, 688-711.

Python Software Foundation. 2020. *Python Language Reference, version 3.9*. Available at http://www.python.org, accessed April 18, 2022.

R Core Team. 2020. *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

SA Innove. 2018. *Inglise keele riigieksami eristuskiri*. Available at https://www.innove.ee/wp-content/uploads/2018/10/Inglise-keele-RE-eristuskiri-2018.pdf, accessed May 20, 2022.

TÜ Senate. 2020. *Admission Rules at the First and Second Level of Higher Education in the 2020/2021 Academic year*. Available at https://adr.ut.ee/?page=pub_view_dynobj&pid=67670123&tid=&desktop=57835&u=&r_url=%2F%3Fpage%3Dpub_view_dynobj%26pid%3D68502154%26tid%3D58359%26u%3D20220520201201, accessed May 20, 2022.

Wolfe-Quintero, Kate, Shunji Inagaki, and Hae-Young Kim. 1998. *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii Press.

Zheng, Cui, Tae-Ja Park. 2013. An Analysis of Errors in English Writing Made by Chinese and Korean University Students. *Theory & Practice in Language Studies*, 3: 8, 1342-1351.

APPENDIX

All L2SCA counted and calculated syntactic complexity measures for BA and MA theses. The numbers in the brackets for MA theses are the second smallest results since the shortest thesis was an outlier.

Type of Thesis	*Index*	*Average*	*Maximum*	*Minimum*
BA	Words	6743.63	11732	3541
	Sentences	284.86	618	161
	Clauses	600.04	1119	333
	Verb Phrases	805.63	1465	436
	T-units	322	674	181
	Dependent Clauses	264.51	520	104
	Complex T-units	176.37	363	89
	Coordinate Phrases	186.49	400	64
	Complex Nominals	861.23	1694	467
MA	Words	13600.75	26632	1933 (7698)
	Sentences	579.84	1197	97 (315)
	Clauses	1195.07	2532	141 (590)
	Verb Phrases	1697.11	3282	217 (850)
	T-units	655.16	1396	100 (351)
	Dependent Clauses	510.12	1065	38 (179)
	Complex T-units	341.82	759	32 (143)
	Coordinate Phrases	403.93	800	73 (197)
	Complex Nominals	1747.28	3429	267 (937)

Type of Thesis	*Index*	*Average*	*Standard Deviation*
BA	Mean length of sentence	24.23	4.76
	Mean length of T-unit	21.38	3.87
	Mean length of clause	11.41	1.53
	Clauses per sentence	2.13	0.32
	Verb phrases per T-unit	2.52	0.37
	Clauses per T-unit	1.87	0.22
	Dependent clauses per clause	0.44	0.06
	Dependent clauses per T-unit	0.83	0.22
	T-units per sentence	1.13	0.06
	Complex T-units per T-unit	0.55	0.09
	Coordinate phrases per T-unit	0.59	0.18
	Coordinate phrases per clause	0.32	0.09
	Complex nominals per T-unit	2.75	0.65

	Complex nominals per clause	1.47	0.3
MA	Mean length of sentence	23.83	2.99
	Mean length of T-unit	21.12	2.42
	Mean length of clause	11.66	1.14
	Clauses per sentence	2.05	0.26
	Verb phrases per T-unit	2.61	0.29
	Clauses per T-unit	1.82	20.19
	Dependent clauses per clause	0.42	0.06
	Dependent clauses per T-unit	0.77	0.18
	T-units per sentence	1.13	0.06
	Complex T-units per T-unit	0.52	0.08
	Coordinate phrases per T-unit	0.64	0.14
	Coordinate phrases per clause	0.35	0.09
	Complex nominals per T-unit	2.73	0.42
	Complex nominals per clause	1.51	0.21

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Kadi Haamer

Similarities and Differences in the Syntactic Complexity of Bachelor's and Master's Thesis: A Comparative Study Using L2SCA

Sarnasused ja erinevused bakalaureuse- ja magistritööde süntaktilises keerukuses: võrdlev uurimus kasutades L2SCAd

bakalaureusetöö

2022

Lehekülgede arv: 33

Annotatsioon:

Käesoleva bakalaureusetöö eesmärk on võrrelda erinevate Tartu ülikooli anglistika osakonna õppeastmete õpilaste kirjalikku inglise keele keelekasutust süntaktilise keerukuse kaudu. Töö sissejuhatuses arutletakse lühidalt õpilaste keeletaseme ja lõputööde olulisuse üle ning seletatakse lahti arvutiprogrammide ja õppijakeele mõiste. Esimene sisuline sektsioon vaatab varasemaid uurimusi sarnastel teemadel ja seletab lahti, miks selleks analüüsiks valiti L2SCA arvutiprogramm. Samuti räägitakse lühidalt, milliseid süntaktilise keerukuse indekse töös analüüsitakse. Teine sisuline sektsioon hõlmab empiirilist osa. Esmalt seletatakse tööde puhastamise protsessi ja selleks loodud programmi kasutamise eeliseid ja puudusi. Seejärel seletatakse lahti L2SCA programmi kasutamise protsess ja kuidas arvutati valitud indekse. Järgmised alamsektsioonid analüüsivad tulemusi, võrdlevad neid varasemate uurimustega ja toovad välja võimalikke põhjendusi. Viimaks pakutakse välja teemasid edaspidisteks uurimusteks.

Esmalt teisendati magistritööd TXT formaati ja loodi programm, mis "puhastab" tööd kõigest, mis ei kuulu õppijakeele alla – sealhulgas tsitaadid, viited, lehekülgede numbrid ja tabelid. Seejärel lisati magistritööd varasemalt töödeldud bakalaureusetööde korpusesse. Kasutades L2SCA arvutiprogrammi, arvutati välja kuus lõputööde süntaktilise keerukuse indeksi. Kõigi nende kohta arvutati keskmised tulemused ja standardhälve nii bakalaureuse- kui magistritööde kohta. Lisaks arvutati välja sõnade ja lausete keskmised, madalaimad ja kõrgeimad tulemused. Analüüsi osas võrreldi arvutatud süntaktilise keerukuse tulemusi bakalaureuse- ja magristöödes. Samuti võrreldi tulemusi eelnevates uurimustes kogutud põliste inglise keele kõnelejate tulemustega ja teistest keeletaustadest kogutud tulemustega.

Selles töös uuriti süntaktilisi sarnasusi ja erinevusi bakalaureuse- ja magistritööde vahel Tartu Ülikooli Anglistika osakonnas. Failid puhastati järgides õppijakeele põhimõtteid ning nende lauseehitust võrreldi L2SCA programmi tulemuste abil. Võrdluses leiti, et nii magistri- kui bakalaureusetööde puhul oli näha suuri erinevusi pikkustes. Erinevalt varasematele ootustele, olid magistritööde tulemused mitmel juhul madalamad kui

bakalaureusetöödel. Kui Eesti õpilaste tulemusi teiste keeletaustadega võrreldi, seal hulgas inglise keel emakeelena rääkijatega, tuli välja, et Eesti lõputööde tulemused olid peaaegu kõikides süntaktilise keerukuse indeksite puhul kõrgemad kui teistel. Need tulemused viitavad sellele, et Eestlased kipuvad inglise keeles kirjutades kasutama liiga komplekssseid ja pikki lauseid. Selle järelduse kinnitamiseks peab tulevikus uurima lähemalt eestlaste kirjakeelt võrreldes inglise keele või teiste keeletaustade kõnelejatega.

Märksõnad: L2SCA, süntaktiline keerukus, õppijakeel, õppijakeele korpus, bakalaureusetööd, magistritööd, Anglistika osakond, Tartu Ülikool, võrdlev uurimus.

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kadi Haamer,

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

Similarities and Differences in the Syntactic Complexity of Bachelor's and Master's Thesis: A Comparative Study Using L2SCA,

mille juhendaja on Jane Klavan,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Digitaalne allkiri
Kadi Haamer

Tartus, 24.05.2022

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.

Digitaalne allkiri
Kadi Haamer

Tartus, 24.05.2022

Lõputöö on lubatud kaitsmisele.

Jane Klavan

Tartus, 24.05.2022