UNIVERSITY OF TURKU

Word frequencies and lexical richness in L2 writing: a study on Finnish learners of English from different levels

Ilari Ratamo

Master's thesis

School of Language and Translation Studies

Degree Programme in Language Learning and Teaching, Department of English

Faculty of Humanities

University of Turku

March 2024

Master's Thesis
School of Languages and Translation Studies / Faculty of Humanities
Ilari Ratamo
Word frequencies and lexical richness in L2 writing: a study on Finnish learners of English from different levels
57 pages, 6 pp. appendices

In this MA thesis three different levels of Finnish learners of English were studied to examine if their differences as English learners could be seen in different lexical measures. These measures concentrated on vocabulary depth and lexical richness. The first aim was to find out how proficiency affects the frequencies of words used in written text. In addition, also academic language use was tracked. The second aim of the study is to find out to what degree does lexical variation and lexical density reveal differences in learners of English.

The groups were comprehensive school, high school and university students. Essays were collected from each group 54, 42 and 51, respectively. All essays were prepared by for example correcting spelling mistakes, in order to make them work with vocabulary analysers that were used to study the data. Two different analysers were used, one for word frequencies, EAP, and one for lexical variation and density, VP-Classic.

The results of the study suggest that word frequencies do differ between proficiency levels and can be used as a predictor to vocabulary development. As the proficiency improves, the more likely the learner is to use less common language. As for lexical variation and density, they did not seem to reveal differences between proficiency groups. Out of six different relations between the groups across these two variables, only one was significantly different and all the results were extremely similar.

In future studies essay topic selection should be thought through with serious caution and the raw data should be prepared with great care to ensure the validity of the results.

Key words: word frequency, lexical density, lexical variation, vocabulary, vocabulary analyser, English as a foreign language, vocabulary depth

Table of contents

List of tables

List of figures

List of abbreviations

NGSL	New General Service List
GSL	General Service List
NAWL	New Academic Word List
AWL	Academic Word List
TTR	Type-token ratio
VP-Classic	Vocabprofile Classic
LFP	Lexical frequency profile

1 Introduction

Vocabulary plays a crucial role in language learning. This can be understood without greater knowledge of how languages work by just having a moment of self-reflection and coming to the realisation that the thoughts one thinks are, in one way or another, realised as words part of that individual's vocabulary. Words are needed to understand and to be understood. In language research this is a commonly accepted fact, which can be illustrated by the several researchers whose primary field of study is vocabulary, Nation, Schmitt, Read, Laufer, Meara to mention a few. It can also be illustrated by statements made about vocabulary by well-established researchers like Gass and Selinker, who stated that vocabulary might be the most important component for language learners (Gass & Selinker 1994: 270).

This has led language teachers and researchers to the search of the optimal way of learning vocabulary. In Finland this has become an important task because of ever growing globalisation, which has caused second language mastery to be in increasing demand. English as a second language has an especially strong emphasis in this subject because of its growing position as a lingua franca among western countries. As one's vocabulary is ever growing and never really complete, the task of efficient vocabulary learning is present in all levels of Finnish education system, starting from pre-school, going through comprehensive school, continuing with high school, and ending at university level.

Vocabulary knowledge has been studied in several levels of the Finnish education system, but the studies usually have concentrated on one or two of the aforementioned levels like the MA thesis done by Ala-Akkala (2010), which concentrated on comprehensive school and upper-secondary school students. In turn, Jaatinen & Mankkinen 1993 and Makkonen 2008, concentrated in university students. The present study was made to fill this gap by being a comparative study between three levels of education, comprehensive school, high school, and university level. Furthermore, the present study concentrates on lexical richness rather than vocabulary size.

The present study has two research questions first of which concentrates on word frequency levels and academic vocabulary as it asks: How do the educational levels affect the percentual usage of word frequency levels and academic vocabulary used in writing? The second research question in turn concentrates on lexical densities and type-token ratios by asking: To

what extent do lexical density and lexical variation reveal differences between L2 writers from different educational levels?

These questions are answered by investigating percentual usage of different word frequency levels and academic vocabulary as well as calculating type-token ratios and lexical density in essays written by students of the three educational levels. The percentual usage of different frequency levels and academic vocabulary was analysed by using a vocabulary profiler provided by EAP Foundation, which uses the New General Service List and the New Academic Word List as its corpora. Type-token ratios and lexical densities, on the other hand, were calculated by using a different vocabulary analyser called VP-Classic. The essays used in the present study were collected already in 2014 by the University of Turku, which allowed to concentrate on analysing them rather than acquiring them. Around fifty different essays were collected from each level of education. The comprehensive school students were ninth graders, high schoolers were in their final year and university students were all English students. Three different topics were presented to the participants of which they could freely choose from. The present study was conducted by modifying these essays to be compatible with the vocabulary analysers and then ran through the analysers after which statistical analysis was applied. The study is quantitative in nature since the results are all numerical.

The thesis continues from here with chapter 2, which includes an overview of relevant theories, terms and other aspects of vocabulary learning and research. Starting with the definition of terms like “word” and “vocabulary” as well as the latter’s different categorization options. Furthermore, chapter 2 includes what it means to know a word, why vocabulary knowledge is important and relevant terms including but not limited to “academic vocabulary”, “lexical richness” and “type-token ratio”. Chapter 3 explains the methods used to conduct the present study such as how the data were collected, what vocabulary analysers are which type of them were used in the study and how the data was prepared. Chapter 4 is dedicated to the results of the present study, which are gone through in the order of the research questions. In the second to last chapter, chapter 5, the findings drawn from the results of the study are discussed as well as the limitations of the it. Lastly, chapter 6, concludes the thesis and summarizes the thesis as a whole. Furthermore, in this chapter suggestions for future research regarding the subject of the thesis are presented.

2 Theoretical background

In this chapter I shall introduce the basic theories, concepts and terms as well as explain them in relevance to the present study. I will start by defining the most important term which is the term "word" and different ways of looking at this term. Second, I will explain what vocabulary is and several of its dimensions and what is meant by vocabulary knowledge. Connected to the previous I will explain what it means to know a word. Continuing from there I will introduce different ways of categorizing vocabulary. Next, I will introduce a couple of more relevant terms and give a glimpse of why vocabulary knowledge is important.

2.1 Defining the term "word"

The definition that most of the population is the most familiar with would probably be the orthographic definition of *word.* This means that *word* is a sequence of letters that is left between two spaces, a space and a full stop or perhaps a space and a comma among other possible combinations (Carter 1998: 4-5). Using this definition, the following sentence "Kittens aren't kittens." would have three items but when talking about items that are mentioned multiple times this type of definition poses a problem since using a word twice does not mean that one would know two words. Also, contractions like "aren't" are also problematic since with the present definition they would also be considered as just one word.

Contractions are two different words compounded into one so this is a major reason to not use this definition as a basis for this type of academic research. But even if we deconstructed the "aren't" into "are not" in our example sentence we would still run into a problem with repeated items. I doubt that many would try to argue that knowing the phrase "Like baby, baby, baby, no" from Justin Bieber's song "Baby" would equate to knowing five different words. To make it more clear why this is the case we can think of a literal baby who just has learnt his or hers first word, which they are now repeating dozen times per hour. Now, if we calculate how many times the baby uses their word per day, we get an amount of 144, if we assume the baby is awake twelve hours per day. Even though this baby has used this word 144 times in the last twelve hours it would be extremely hard to argue for this baby knowing 144 different words. To solve these types of issues researchers have come up with terms *type* and *token*.

Tokens and types are especially useful if one's goal is to count words. This is because tokens use the orthographical definition of a word, which means in the context of counting them that each and every string of letters between spaces or punctuation marks are counted as one word (e.g. Nation 2001: 7, Schmitt 2010: 188). But, if this was the only way to count words, we would have to say that knowing "Like baby, baby, baby, no" and "Kittens really are cute though." would be equal when testing how many words one knows since both of these phrases have five tokens.

The term type allows to work around this problem because when using types to count words if we run into the same string of letters twice in one data set, we do not count that string of letters again (ibid.) as we would have done when using tokens. Using types to count the words in "Like baby, baby, baby, no" our results would show that this sentence contains three words but when using them to count the words in "Kittens really are cute though." we would have five words.

Type and token still do not offer enough coverage in the world of word definition because of for example inflected forms. In the sentences "I studied yesterday. When did you study?" we have six different words regardless of which of the previously mentioned definitions we use, even though if we think about it, it used twice the word "study", it was just inflected in its second appearance. A way to combat this problem is the use of *lemmas*. Lemmas see words more like complete abstract lexical units which include several inflected forms of its headword (Nation 2001: 7-8). As in our example "I studied yesterday. When did you study?" every other word would count as their own lemma but as mentioned before "studied" is just an inflected form of "study" and that is why they do not count twice when counting lemmas. All of the inflections that lemmas in English include are the following: plurals, third person singular present tense, past tense, past participle, *-ing*, comparative, superlative and possessive (Bauer & Nation 1993). In addition to the previous, lemmas also include the reduced forms of words like "aren't" and "can't" (Nation 2001: 7-8).

Next there are *word families*, which are a broader way to look at words than lemmas since alongside of the headword also "its inflected form and its closely related derived forms" (Nation 2001: 8) are included in them. This means that if one looked at the word "forgive" all of the following would be included in a singular word family: forgive, forgiven, forgivable, forgiveness, unforgiving and unforgivable. The first two of them are examples of its inflected forms and the rest are examples of its derived forms, which were formed by adding prefixes

and suffixes to them. According to Nation and Beglar (2007) when measuring vocabulary size, word families should be the preferred way of counting words because learners with more than minimal proficiency in the target language are able to control word building devices (such as affixes) to derive many different words with just one affix and multiple headwords (Nation & Beglar 2007: 10). These kinds of learners are also "able to see that there is both a formal and a meaning relationship between regularly affixed members of a word family." (ibid.).

Lastly on this topic, there are *flemmas*, which are a term going between lemma and word family, even though they resemble lemmas more than word families. The main difference between lemmas and flemmas is that flemmas do not differentiate between different parts of speech (Stoeckel, et al. 2020). This means, for example, that the verb "developed", and the adjective "developed" would count as two items if one chooses to use lemmas when counting words but when using flemmas these items would count as one. Why flemmas are important in the case of the present study is that the word lists that are used as a corpus uses flemmas to differentiate between what is a word and what is not.

2.2 Definition of vocabulary

Now that we have established what a word is we can move on to what vocabulary is since if one took all the words out of a vocabulary there would be nothing left. This is because *vocabulary* as a term refers to a component of language that involves words and their meanings (Saville-Troike 2006: 191). This might be a bit simplified way of defining vocabulary as a whole since a more common viewpoint is that vocabulary is a multi-dimensional construct rather than a single dimension. In the following section I shall differentiate between two main dimensions of vocabulary since the distinction between them is important to the present study. Explaining the differences between the two dimensions will also give additional information about what it means to know a word, a subject that will also be explained in more detail. In addition, I will explain relevant concepts and terms such as type-token ratio and lexical richness.

The main distinction in vocabulary knowledge is usually made between productive and receptive skills. Productive vocabulary knowledge is needed when one attempts to use vocabulary to produce something i.e., writing and speaking (Saville-Troike 2006: 137, 193).

Receptive vocabulary knowledge on the other hand is needed when one attempts to receive this produce that someone else with their productive knowledge has produced which is done via listening and/or reading (ibid.). It could also be said that productive knowledge is used when you try to be understood and receptive knowledge when you are trying to understand. The productive/receptive division is not as clear cut as one would think because even though it would be incredibly convenient for vocabulary studies and teaching there is no real consensus on how these should be accurately defined or whether the distinction is psychologically valid to begin with (Melka 1997). There are two ways of looking at this problem first of which is the simple answer that the distinction between the two knowledge types is clear, and they really are different types of knowledge and the second is that there is a continuum between these two where different levels of knowledge exist (Nation 2001: 24-26).

Imbalance between how much we understand from for example reading and how much we are able to deliver while speaking for example is not unheard of. This might be because several studies have shown that when testing the test subject's receptive and productive vocabulary knowledge size the receptive ones seem to be larger compared to the productive ones (Melka 1997). In addition to this it also seems that learning receptive vocabulary knowledge is also easier than learning productive knowledge (Nation 2001: 32-33). Melka (1997: 99) does however point out that this might be a result of methodological problems in these studies, which leaves a reason to doubt if these two knowledge types were reliably tested. Melka continues that vocabulary tests should not be classified as productive or receptive knowledge tests because their definitions are not clear (ibid.). Even though the present study does not include a vocabulary test it should still be mentioned that when using the terms productive and receptive vocabulary knowledge, of which the first has even more emphasis because of the nature of the present study's data, they are not used as counterparts but rather to refer to the fact that the raw data used in this study had been gathered in written form.

In addition to the receptive-productive idea, a reoccurring viewpoint to look at vocabulary knowledge is that even if one knows a lot of words it does not mean that they knew them well. To clarify this viewpoint terms vocabulary breadth and vocabulary depth were developed. *Vocabulary breadth* (also known as vocabulary size) refers to the number of words one knows regardless of how well one knows them (Qian 1999, 283). Testing this type of vocabulary knowledge can be relatively easy and it is commonly done via checklist-tests (Meara 2010, 15). These types of checklist-tests are commonly used in schools when conducting a word test of textbook chapter's words. *Vocabulary depth* on the other hand,

broadly speaking, refers to how well the words are known or the quality of the knowledge (ibid.). How to define "quality of vocabulary knowledge" is not an easy task but Nation (2001) has given a framework, which will be gone over in the next section. This framework consist of nine aspects and Nation proposes that all of these aspects together illustrate what it really means to know a word (Nation 2001, 26). Returning to vocabulary depth and how to test it, Nation's (2001) model is comprehensive but because of its comprehensiveness it can be quite a challenge to design a test that would involve all of nine aspects. If one tested all of these aspects it would be a very time-consuming way of testing word knowledge (Meara 1996, 44). There have been attempts at this though, but they have not come up with anything too significant (Read 2000, 178-180).

On the matter of quality of one's vocabulary there is also the term of lexical richness. Lexical richness can be looked at from different angles and according to Read (2000) there are three of these different angles: lexical variation, lexical sophistication and lexical density. Starting with lexical variation, which is meant to give an estimate of how varied vocabulary is used in a given text (Read 2000, 200). This can be studied by analysing type-token ratios (ibid.). Type-token ratio means how many types are in a given text compared to the number of tokens in the text. For example, in the phrase "Like baby, baby, baby no" there are five tokens and three types as gone over earlier and dividing the number of types (3) by the number of tokens (5) we get the number 0.6, which would be the type-token ratio of "Like baby, baby, baby no". Lexical variation and type-token ratio are used as synonyms throughout this study. Lexical sophistication on the other hand can be studied by the checking how many low-frequency words are in a given text (Read, 2000). The usage of these type of words allows the writer to express themselves in a precise and sophisticated way (ibid.). Lastly, when talking about lexical density we mean the percentage of content words in a given text (ibid.). This is calculated by counting the content words and dividing the number by the total amount of tokens in the text. In addition, writing quality can also be measured by the number of errors found in a text (ibid.).

2.3 How to know a word

When talking about vocabulary knowledge we also have to define what it means to know a word. This question is not as easy to answer as one might think since most of us have taken part in some kind of word test in our lifetime where we have been tested whether or not we

know a word and those tests mostly have been quite straight forward and simple. These types of checklist tests that are usually done in schools have been designed to test if one knows the meaning of a word, which is the most important part of vocabulary knowledge. But sometimes this is not sufficient since there are much more to knowing a word than just being able to define its meaning like how to use a certain word in relation to other words for example. Nation (2001, 27) has divided word knowledge into three different categories shown in the table below.

Table 1. What is involved in knowing a word (Nation 2001, 27). (Numbers in parentheses are added by me to allow easier following of the example used below.)

Form	Spoken (1.1)	R	What does the word sound like?
		P	How is the word pronounced?
	Written (1.2)	R	What does the word look like?
		P	How is the word written and spelled?
	Word parts (1.3)	R	What parts are recognizable in this word?
		P	What word parts are needed to express the meaning?
Meaning	Form and meaning (2.1)	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concept and referents (2.2)	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations (2.3)	R	What other words does this make us think of?
		P	What other words could be use instead of this one?
Use	Grammatical functions (3.1)	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	Collocations (3.2)	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use (register, frequency…) (3.3)	R	Where, when and how often would we expect to meet this word?
		P	Where, when and how often can we use this word?

Note: In column 3, R = receptive knowledge, P = productive knowledge

Here we have three categories: form, meaning and use and connected to each of these three there are three subsections, which specify what type of features are associated with the three main categories. On each of the subsections there are two questions which demonstrate what

questions one should theoretically be able to answer, if one truly knows a word. Furthermore, on the left side of these questions there is also a letter R or a letter P indicating whether the question is asked from the receptive vocabulary's (R) or the productive vocabulary's (P) point of view. To demonstrate how to use this table I shall introduce you to Nation's (2001) example of what it means to know the word "underdeveloped" using the table above.

To know the form of "underdeveloped" from receptive point of view is to be able to recognise the word when one hears or reads it (1.1 & 1.2). One also has to understand that the word is made up of parts "under-", "-develop-" and "-ed" as well as how these parts affect the word's meaning (1.3). To understand the meaning of the word one has to know that "underdeveloped" means something in particular as well as what the word means in the context it appears in (2.1). One also must understand the concept behind the word to be able to understand what it means in variety of contexts (2.2). Furthermore, one is needed to be aware of the word being related to words like "overdeveloped", "backward" and "challenged" (2.3). If one wants to know the use of "underdeveloped" one must be able to recognise when the word is used correctly in a sentence as well as recognise that typical collocations with the word are for example territories and areas (3.1 & 3.2). It also has to be understood that the word is not a particularly uncommon word nor a pejorative word (3.3) . Moving on to knowing the form of our example word from the productive point of view one first of all has to be able to pronounce and write it correctly including the ability to construct the word of its parts and their appropriate forms (1.1 & 1.2 & 1.3). If one understands the meaning of "underdeveloped" one is able to produce it in a manner that expresses its meaning and is also able to do this in various contexts to be able to use the range of meanings it possesses (2.1 & 2.2). One also must be able to produce synonyms and opposites for it (2.3.). Lastly, to understand the use of our example from the productive point of view one has to be able to use it correctly in an original sentence, produce words commonly occurring with it as well as use it appropriately according to the formality of the situation it is used in, which might mean replacing it by another word (3.1. & 3.2. & 3.3). (Nation 2001, 26-28)

Using this type of solving technique it is possible to pick apart what it means to know any word.

2.4 Different ways of categorizing vocabulary

There are multiple ways to categorize vocabulary but in relevance of the present study there are four that will be explained: content and grammatical words, K-bands and the Academic Word List (AWL). The first two are their own pair because they all deal with what the words in their own category mean and how they behave in language as compared to the latter two which deal more with the question of what type of words belong to what K-band or what type of words belong to the AWL. In this following section I will go through all of the aforementioned and explain what they mean starting with content and grammatical words. In addition to this I will also explain what are the (New) General Service List and the (New) Academic Word List.

The easiest way of identifying whether or not a word is a content or a grammatical word is to look at its meaning. If it belongs to the latter group, it will have close to no meaning or no meaning at all when looked at independently because it mostly plays, like the name suggests, a grammatical role in language (Singleton 1999, 11). Examples of grammatical words would be in, the, and shall. These types of words are also sometimes referred to as function words or structure words and they belong to a closed and finite category of words, which means that it is extremely rare for them to get new members (Greenbaum & Quirk 1990, 15-16). Comparing this information to content words we see differences in all aspects. While grammatical words have basically no meaning content words have substantial meaning even without any context (Singleton 1999, 11). Examples of these would be house, fly, and yellow. Content words belong to open word classes, which means that new words are coined often, compared to closed word classes, to join these open classes and also words drop out of these classes as language develops (Greenbaum & Quirk 1990, 16).

When talking about K-bands we are referring to a term which is used when measuring word frequencies i.e., how often or how seldom words are used. To give a little background of word frequencies, especially in English, a good example are the words that are used the very most since they are a major part of almost any English text or speech. Starting from the very top the most used word in English is “the” and it alone accounts for 7 % of all words in a given text or speech (Nation & Waring 1997, 8). Continuing to the ten most commonly used words they in turn account for about 25 % of a given piece of text or speech (Laufer & Nation 1999, 35). This of course does not mean that if one learnt just the ten most used words of English that one would be able to understand 25% of any text they would read. The most frequent words

account for a disproportionate number of items of texts hence it is useful to look at vocabulary as a series of levels divided by the frequency of the vocabulary's words (Laufer & Nation 1999, 35). These levels are also known as K-bands. Each level or each band consists of a thousand words of which the first is the thousand words that are most commonly used or as it would be in K-bands, K-1. K-2 then refers to the next thousand words so the from the 1001st most frequent word to the 2000th most frequent word. These levels go on in this manner the further into K-bands one dives. Below in Table 1 there are the six first K-bands of English according to the Brown corpus (Francis & Kučera 1982), which consists of over 1 000 000 tokens from different kinds of types and registers.

Table 2. Frequency band and percentage of text coverage

Frequency level	**Cumulative coverage (%)**	**Coverage (%)**
1st 1000	72.0	72.0
2nd 1000	79.7	7.7
3rd 1000	84.0	4.3
4th 1000	86.8	2.8
5th 1000	88.7	1.9
6th 1000	89.9	1.2

As seen in Table 1 there are major fluctuations between the coverage of these different K-bands. Coverage in this case means how large percentual part of all the words of the data set is part of one k-band. The largest being, by a considerable margin, the first having a coverage of 72%. Comparing this percentage to the coverage of the second k-band (7.7 %) one can see a difference of 64.3 %. These two together account for almost 80 % of all words of the Brown corpus, which in itself already is a huge portion of it. Going down from them the coverage by K-band becomes smaller and smaller and by the 6th band one can see that it covers only 1.2 % of the corpus. Because of this the first two bands are often called *high-frequency words* while the others are called *low-frequency words* (Nation & Hwang 1995).

A famous and widely used list of high-frequency words of English is the General Service List (GSL) put together by West in 1953, which consists of about 2 000 words of which around 160 are grammatical words and the rest are content words that appear regularly and in many kinds of registers of English. Since West's list was compiled in 1953, recently a more up to date general service list was compiled called the New General Service List (NGSL). The list was put together by the team of New General Service List Project, which consists of Charles Browne, Brent Culligan and Joseph Phillips. The NGSL was published in 2013 and it was built on the original GSL by expanding, modernizing and improving upon their research (Browne, Culligan and Phillips 2023). This new list was based on a subset of 273 million words of the Cambridge English Corpus and has been updated in 2016 and 2023 (New General Service List Project 2023). The newest version of the NGSL, NGSL 1.2, consists of 2809 words and with these 2809 words it covers an average of 92 % of most texts of general English (ibid.). In some situations, it covers even larger amount for example it covers 93 % of the vocabulary used in the Harry Potter series, 94 % of TOEIC (Test of English for International Communication) exams and 95 % for many TV-shows, like a show called "Friends" for example (ibid.). Another notable difference between the NGSL and the GSL, in addition to the previously mentioned, is that NGSL does not use word families but flemmas in its list (ibid.).

Nation (2001) saw the need to distinguish further than just high and low-frequency words and he proposed the idea of academic words and technical words of which in the light of this study just the first one is relevant. There have been other lists that propose their own version of academic words mainly the University Word List (UWL) by Xue and Nation (1984) but in the light of the present study I will explain more thoroughly the Academic Word List (AWL), which was put together by Coxhead in 2000. The basic idea behind the AWL is to have a list of words that appear frequently in academic texts but are outside the high-frequency words. This resulted in a list consisting of 570 words which cover around 10 % of all tokens in Academic Corpus, which itself consists of over 400 academic texts from various fields (Coxhead 2000). When comparing this 10 % to Table 1 and looking at the second K-band we can see that this list of just 570 words covers more of the words in academic texts than the second most frequent 1 000 words. There are differing results between corpora, but the coverage of the AWL is of such significant that the point still stands. From this we can conclude that if there was a person knowing only the most frequent 1 000 words of English

and the person would only want to read academic texts, they would be better of learning the list of 570 words in the AWL rather than the second most frequent 1 000 words of English.

Just like in the case of the GSL the AWL also was seen to be in need of updating and thus, the New Academic Word List (NAWL) was compiled. The NAWL is a word list of 957 words, which the creators of this list, the team of the New General Service List Project, claim to be the most important words for general academic English (New General Service List Project 2023). They say that it is an update and replacement for Coxhead's AWL and it was needed to make it compatible with the NGSL and because they also hoped to attain a higher coverage of its target texts (ibid.). They claim that the NAWL together with the NGSL give a larger coverage of 5 % compared to the AWL and the GSL put together (ibid.).

2.5 Vocabulary in second language learning

Vocabulary consist of words and words is one of if not the main way of communicating with the world and everything in it. Without words communication in the world of today would be significantly more difficult. Goulden, Nation and Read calculated in 1990 that the *Webster's Third International Dictionary* contains 114 000 words while excluding proper names. They also did another calculation where they also excluded compound words, abbreviations, alternative spellings, archaic words and dialect words and the result was still around 54 000 words. These kinds of numbers of course are way out of the regular first language speaker's range since it has been estimated that a good rule of thumb is that native speakers learn about 1000 words per year (Nation & Waring 1997, 7). This estimation is valid until the speaker reaches approximately a vocabulary of 20 000 words meaning that by the time speaker reaches the age of 20, they already have learned somewhere between 17 000 and 20 000 words (ibid.). There have been studies that give differing results to Nation and Waring. Takala (1989, 7) did a study where they came to the conclusion that a 6-year-old would know about 6 000 head words, high school students approximately 150 000 head words and college students would go as high as 200 000 head words. To note of this study is that the term "head word" was left undefined.

We now have given estimations about native speaker's vocabulary and the numbers might seem huge but L2 learners have different kinds of vocabularies. Even though there have been studies that show that L2 learners learn new words at the same rate as L1 speakers this does

not seem to be the case in most of the typical scenarios since a multitude of adults with English as their L2 language seem to know less than 5 000 words (Nation & Waring 1997, 7-8). There have been a couple of studies regarding the vocabulary of Finnish university students of English, which I will go through next. In these studies, different kinds of vocabulary tests were used which will be gone over at a later point. Jaatinen and Mankkinen ([1992] 1993) conducted a study using dictionary-based vocabulary test where they had two groups from the University of Jyväskylä and compared their receptive vocabulary sizes. The two groups were BA students and MA students, 52 participants and 37 participants respectively. The results showed that BA students knew around 17 000 words while the MA students knew around 19 500 words. Combining these figures, we get to the conclusion that according to this study the average university student of English knows about 18 100 words.

Another study that used Finnish university students as participants is Makkonen's MA thesis from 2008 where they investigated the vocabulary of students at the University of Turku. This time the two groups were divided into first year students and students who had been studying English for more than a year. It should be noted that Makkonen used Schmitt's revised version (2001) of the Vocabulary Levels Test, which significantly differs from dictionary-based vocabulary test as gone over earlier in this thesis. Makkonen (2008, 52) herself said that rather than providing a clear single figure of one's vocabulary size the Vocabulary Language Test lays out a profile of the learner's vocabulary. The results were that the average score was 139.31 correct words out of 150 words, and that there was no significant difference between the two groups (Makkonen 2008, 65). These differing results comparing to Jaatinen's and Mankkinen's study discussed previously might be, as Makkonen points out (2008, 62), because of a ceiling effect, in other words, the test just was not hard enough to differentiate between these two groups. Though it could also be seen from the view point that receptive vocabulary knowledge does not improve that much during university studies. Laufer (1994) conducted a study that was interested more in the productive side of vocabulary knowledge where they studied university students of English in Israel. In this study the participants were asked to write free compositions at three different points of time, first of which was collected from the entrance exams, second at the end of first semester and third at the end of the second semester. The participants were motivated to write as well as they were able to because the first composition could be the deciding factor on whether or not they would be getting into university and because the two later compositions were given the weight of a course exam. The result of this study was that there was a significant change in vocabulary richness when

looking at it in the terms of how many words beyond the most common 2000 words the participants used. These changes were present in both of the latter compositions (Laufer, 1994).

For a considerable period of time the importance of vocabulary was left outside the spotlight in L2 teaching while things like grammar were given more time in it (Laufer & Nation 2012, 163). These days the spotlight has moved clearly to the direction of vocabulary since the linguists nowadays have a consensus among themselves that L2 vocabulary knowledge is an essential part of being proficient in an L2 (Henriksen 2006). Laufer (1992, 101) even suggested that vocabulary knowledge presents the biggest obstacle while trying to advance in L2 proficiency. In addition to this there have been a number of studies showing that L2 learners with not so large vocabularies perform weaker in several L2 skills compared to learners with larger vocabularies (e.g. Meara 1996). Furthermore, Schmitt (2010, 4) suggests that vocabulary knowledge clearly plays a role in almost all fields of L2 competence.

Vocabulary obviously plays a major role in all four (writing, speaking, reading and listening) areas of L2 competence but it has been suggested that the need for vocabulary knowledge might be increased in writing and reading compared to speaking and listening (e.g., Nation 2001, 125). However, Nation (2001, 125) himself has pointed out that the difference might just be due to the fact that usually more important matters are gone over in writing, which results in topics that might not be that familiar compared to topics that are gone over while speaking. Several lexical studies have shown a clear correlation between the size and depth of vocabulary size and writing skills of L2 learners (Henriksen & Danelund 2015, 29). Furthermore, vocabulary can be used as a predictor of how a person would perform on written assignments (Henriksen & Danelund 2015, 31). Comparing more proficient L2 learners with less proficient learners it has been found that the former use richer vocabulary in their compositions than the latter (Nation & Meara 2002, 51).

3 Methods

In this section I shall go over and explain the methods how the present study was conducted. This will include what the original data consists of and how it was collected. In addition, since vocabulary profilers were used to analyse the data, I will explain what they are and what type of vocabulary profilers were used in the present study and why. After this, I will explain how the data was prepared to be applicable for the present study.

The data was collected by the University of Turku, not specifically for the present study, but to gather data about different levels of Finnish education system in the hopes of it being useful for future studies like the present study. Data was collected from three different levels of education: comprehensive school, high school and university. The initial collection was conducted already in 2014. The data was collected from several schools and/or institutes on each of the three levels but the present study used data from only one comprehensive school, two high schools and one university. This was a result of one main reason. Finland is a country with two official languages: Finnish and Swedish. Because of this each of the three educational levels also included at least one school with a lot of participants whose mother tongue is not Finnish, and since the present study concentrates on learners with Finnish as their mother tongue some of the data had to be excluded. Regarding the same topic, all of the participants who had not stated that their mother tongue is Finnish were also excluded. Other limitations regarding the use of the data in the present study's framework were the following. As the data was collected in both written and spoken form only the former would be used for the present study since the vocabulary analysers used are only able to handle written text. Consent had to be given by the participant to allow the usage of the data for research purposes. Lastly, if the data did not exceed a total of 50 words in length or if the data given by the participant was an obvious joke, then, it had to be excluded.

These factors provided, in the end the data of the present study consists of 147 essays of which 54 are from comprehensive school level, 42 from high school level and 51 from university level.

In the instructions the participants had three different subjects that they could freely choose of. The subjects were the following: "Computers in my life", "16-year-olds should be allowed to drive a car in Finland!" and "A person I admire".

In comprehensive schools the data was collected as part of a normal 9th grade English class, which lasted 45 minutes. The schools requested that the instructions for the task would be translated into Finnish and to give some key words and questions to make the task more familiar to the participants. Examples of the key words and questions would be "social media, safety, role model" and "How often do you use the internet?", "How does this differ from the current law?", and "Why do you admire him/her?" (freely translated). In addition to the subjects, the instructions also included a guideline of regarding the length of the task and for comprehensive schools it was 80-100 words.

In high school and university level the instructions were identical to each other but differed slightly from the comprehensive school instructions, mainly on two points. Firstly, the guided length was 150-250 words rather than 80-100, and secondly, each of the three subjects had small added instructions about the imaginary target group to which this text would be targeted to. For example, in the subject regarding computers the participants were instructed to introduce the topic to a group of senior citizens. The participants were given 45 minutes to complete their task but, at least for the university students, in practice 20 minutes was enough for 95% of them.

Regarding research ethics of the data collection in this thesis, everything was done anonymously meaning that attributes of the participants such as, name, sex and religion were not collected and such, did not possess a possibility to affect the present study. Moreover, this applies to geographical matters such as where the schools were located or which schools the participants attended. Continuing with research ethics but from the essays' point of view, all proper nouns such as names and places were changed by the personnel of the University of Turku as they transcribed the essays for future use.

3.1 Vocabulary profilers

There are several different kinds of vocabulary profilers that could be used to evaluate one's vocabulary knowledge but in the case of the present study one called *VP-Classic* (short for Vocabprofile Classic) and another one without an explicit name provided by EAP Foundation were best suited for the job. Henceforth, the latter will be referred as EAP for sake of clarity unless mentioned otherwise. Both of these vocabulary profilers are browser-based programs that are free to use and can be found on online, VP-Classic from lextutor.com and EAP on

eapfoundation.com. The programs are based on Laufer & Nation's *Lexical Frequency Profiles* (LFP), which is used to analyse one's vocabulary breadth. The idea behind LFP is to divide the words of a text to four different frequency categories: the first two K-bands, words found in Academic Word List and words not found in any of the previous three, the last one being called off-list words. The LFP shows how many of the words in the analysed text belong to each of the four categories in percentages. As an example, let us imagine a text consisting of 200 words. From the 200 words 160 belong to the first K-bands, 20 to the second K-band, 10 to AWL and 10 to off-list words. Using these numbers, we would end with results looking like this: 80%-10%-5-%-5% and this would be called the LFP of the analysed text. LFP has been shown to correlate with other lexical measures and to discriminate between different proficiency level learners as well as being relatively stable when comparing two pieces written by the same learner (Laufer & Nation 1995). LFP was also used in a study which analysed essays which were used as applications to a Quebec university's TESL (Teaching English as a Second Language) program (Morris & Cobb 2003). Comparing candidates who got in and who did not get in the former used more words belonging to AWL and less of lower frequency words as the latter used more frequent vocabulary.

EAP does all of the above mentioned automatically with one major difference. It does not use AWL and GSL as its corpus but NAWL and NGSL. From the perspective of the results this is crucial because now there are not only two different K-bands but three, K1, K2 and K3. Other than that, the results look exactly the same, they just use different corpora. Additionally, EAP also lists supplemental words, which are written out numbers, and numerically written numbers. It also lists so-called "off-list words" to which all words not found on either corpus will go to. Below are two pictures that demonstrate how example results would look like on EAP.

Data table

Number of words (total and unique) in each level

Level	TOTAL			UNIQUE		
	# words	%	cumul %	# words	%	cumul %
NGSL 1k	40	87%	87%	29	85.3%	85.3%
NGSL 2k	1	2.2%	89.1%	1	2.9%	88.2%
NGSL 3k	2	4.3%	93.5%	1	2.9%	91.2%
Total for NGSL	43	93.5%	93.5%	31	91.2%	91.2%
NAWL	0	0%	93.5%	0	0%	91.2%
Supplemental	0	0%	93.5%	0	0%	91.2%
Off-list	3	6.5%	100%	3	8.8%	100%
Numbers	0	/	/	0	/	/
Totals	46	100%	100%	34	100%	100%

Picture 1. Example results of EAP

Word lists

Sorted by frequency

NGSL 1k words: a, an, and, animal, applied, been, care, center, city, did, for, get, had, have, i, in, interview, it, job, main, my, not, of, take, the, to, was, would, yesterday
NGSL 2k words: unfortunately
NGSL 3k words: shelter
NAWL words: *none*
Supplemental words: *none*
Off-list: ducks, helsinki, residing
Numbers:

Highlighted text

Different frequencies in different colours

I had a job interview yesterday. It was in Helsinki city center and the job I applied for was in an animal shelter. My main job would have been to take care of the ducks residing in the shelter. Unfortunately, I did not get the job."

Picture 2. Example results of EAP

The example text that was fed for the program is: "I had a job interview yesterday. It was in Helsinki city center and the job I applied for was in an animal shelter. My main job would have been to take care of the ducks residing in the shelter. Unfortunately, I did not get the job."

On the left side of picture 1. are all of the aforementioned categories with the respective numbers of words as well as percentages of all tokens that appeared in the example with their respective cumulative percentages as well. On the right on the other hand, are the same information but this time the words were prepared as types rather than tokens.

In picture 2 are again all of the categories with their corresponding colour codes (which can be determined by the user), light blue for K-1, green for K-2, yellow for K-3, dark blue for NAWL (which there were none of in the example prompt), grey for supplemental words (again, none of) and no colour for off-list. All the words are listed to their respective category, sorted by the frequency of the words. Duplicates are not mentioned in these lists. Under the categories is the whole prompt, which is also colour coded so one can see where each word appeared in the prompt.

In picture 2 on the left are each of the five categories in their own colour codes, light blue for K-1, green for K-2, yellow for K-3, dark blue for NAWL (which there were none of in the example prompt) and no colour for off-list.

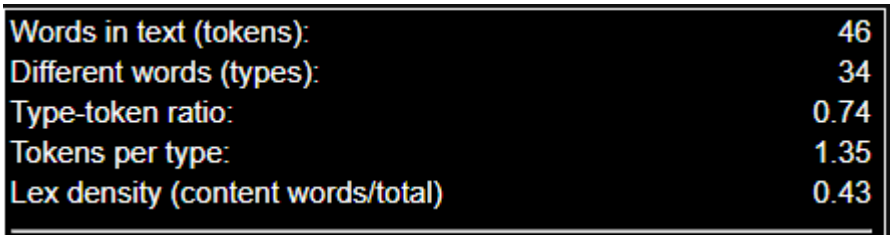
Note that all punctuation marks and capital letters have been removed from the prompt. This is because the program does not comment on how these words were used. It only differentiates between whether it recognises the word or not, but it does not know the context or any other additional information and because of this it only gives information about vocabulary breadth and not vocabulary depth.

Using these information pieces, estimations can be given about the lexical variation of a composition thus also determine the vocabulary knowledge of the writer of the composition in question. Ideal result using these type of vocabulary analysers would show less competent users using more words in the first K-1 band and fewer AWL words and more competent writers using more K-2, K-3 and AWL words. Especially AWL words should see a difference when comparing less and more competent writers as Laufer (1994) found in her study of Israeli learners.

VP-Classic on the other hand is not used to determine the frequencies of words but rather aspects that are not affected by used corpora. EAP and VP-Classic both give the answer to how many tokens and types are in any given text but while using both of these some small differences were noted that lead to the selection of VP-Classic on these kinds of matters. EAP for example does not deal particularly well with anything else but letters and numbers because

for instance the word “o’clock” is interpreted as to different words “o” and “clock” which is not the proper interpretation. VP-Classic on the other hand did not seem to have these kinds of problems so the total number of words was more accurate. In addition to this, VP-Classic also provides type-token ratios, lexical densities and how many tokens were in the text per types.

Below is picture 3 with example results from VP-Classic. These results were gotten using the same example prompt as before.



Words in text (tokens):	46
Different words (types):	34
Type-token ratio:	0.74
Tokens per type:	1.35
Lex density (content words/total)	0.43

Picture 3. Example results from VP-Classic

Here on the very top of the results is the total number of tokens that the vocabulary analysers found in the text, 46 in this case. On the next line is the total number of word types used in the text, 34 in this case. After that comes the type-token ratio, 0.74, followed by tokens per type, which essentially is the same thing as type-token ratio but reversed. Rather than dividing the types by the tokens, the tokens are divided by the types, which in this case comes out to 1.35. Lastly, there is lexical density, which, as explained in the picture as well, is calculated by diving the number of content words with the total number of words (tokens in this case), which comes out to 0.43 in this case.

One other browser-based vocabulary profile program was also trialled for the present study. It was *VP-Compleat*, which can be found on the same website as VP-Classic. The main difference between this and VP-classic is that this one only differentiates between different K-bands and does not take into account AWL. It is worth mentioning that is not limited to K-1 and K-2 but goes up to K-25. Even though this type of information could have been also very interesting to digest it was the AWL, which drew me to VP-Classic and EAP rather than VP-Compleat. AWL has been used in so many other studies (for example Laufer 1994) that the results of this study would not be as comparable with many other studies if AWL would be left out.

3.2 Data preparation

All of the data was prepared equally, meaning that there were specific alterations that had to be done to the data to get as realistic and fair results as possible and these specific alterations were applied to each individual text so that each of them would be useful apart from the terminal limitations discussed before, i.e. consent or mother tongue, for example.

Below are the three main rules that had to be followed when going through each individual answer to make the present study happen:

1. Open up contractions.
2. Fix spelling mistakes.
3. Remove unnecessary items, like emojis or Finnish words.

All of the rules have been created to make the vocabulary analysers used to conduct the present study work properly. Starting with the first rule, EAP does not work very well with contractions like “don’t” and “should’ve”. It counts them as two separate words as “don” and “t” and “should” and “ve”. This most likely is because it is programmed to dismiss all punctuation marks, including apostrophes, and replace them with a space. This results into the two different “words”. To counter this, all contractions were opened up to their original forms, in the case of the two examples, “don’t” and “should’ve” turn into “do not” and “should have”, respectively.

The second rule has to with the off-list words that the vocabulary analyser lists as their own group. As discussed before these are words not found in the analyser’s corpora. Unfortunately, the analysers do not come with a built-in spell-checker, which means that all misspelled words go almost automatically to the off-list group. This was not the desired outcome, so each answer had to be gone through and spell-checked. There were also some spelling “mistakes”, which supposedly were done on purpose. An example of this can be found in an answer from a comprehensive school student regarding computers, which included the following sentence: “That’s bad thing, cause when I wake up Im tired, veeeery tired!!”. Here, the word “very” has three added e-letters in it, which were removed to ensure it ends up in the right category of words. Also, this example has examples of the first rule and looking at the corrected version: “That is bad thing, cause when I wake up I am tired, very

tired!", one can see how they were corrected. If a contraction was missing an apostrophe, it was still opened up. In this example there is also the word "cause", which is the spoken language version of the word "because". This was not corrected due to it not changing categories. Whether it was corrected or not, it still belonged to K1-band.

Regarding the second rule, no words or inflictions were added to any of the texts to make them grammatically more correct. Opening up contractions is different from adding words because contractions indicate that the writer does know the two words the contraction is made of but decided to use the contracted form for whatever reason. This does not apply to the following example sentence: "And when I go London every summer I want to talk my Finland friends." Here, there are no contractions but there are two "to" words missing from the terms "go to" and "talk to". In the corrected version of this sentence nothing was changed. This is because there is no probable reason to think that the writer knew that the two "to" words are missing, ergo, they would not be added to the text. The following sentence, again from the same answer: "But when I was young I use computer also every day, because I love one game in internet, called Barbie!", uses the wrong tense of the word "use" since it should be in past tense. These types of mistakes were not corrected because there was not enough reason to think the writer would know they made a mistake.

As a last note, unfortunately, the second rule has some subjectivity to it. It is not always 100 % certain what is and what is not a spelling mistake since the original writes could be contacted, which left no better option than just me correcting everything as precisely as possible.

Moving on to the third rule, here are two example sentences from the same answer as the earlier ones.

"I think that computer take my sleep time, because if I use computer at night I don't get (uneen)."

"My opinion is that when you use computer you must go to sleep at ten o'clock!! :D"

In the first one the writer has put in parentheses the Finnish word "uneen", which is an inflicted form of "uni" meaning "sleep". This is most likely because the writer did not remember or did not know how to formulate a sentence explaining them falling asleep, so they substituted the term "fall asleep" with "get sleep", by using the Finnish word for "sleep". In cases like these the Finnish word was deleted from the answer. The same goes for the other

example sentence mentioned where it has an emoji ":D" at the end of it to express emotion. Since the vocabulary analysers used in the present study have not incorporated emojis into their corpora, they also were deleted from the answers.

As a last note regarding the preparement of the data, all of the words used in all of the three different subjects the participants could freely choose from belong to the K1-band except for the word "admire", which belongs to the K3-band. Because of the very seldom use of K3-words in the essays as a whole, it was decided that the word admire would be dismissed from the data altogether. This is because each essay that had "A person I admire" as its subject did use the word "admire" in its title. It can also be argued that having a K3-band word in the given subject gives everyone who decided to go with the subject in hand an unfair advantage to use a K3-band word because it was given to them.

4 Results

In the following section I will introduce the results of the present study. The results are divided into three sections. First, I will show what the results look like over-all, without differentiating between groups. The next two parts are divided according to the research questions of the thesis, starting with the first one and continuing with the second one. These two sections also contain comparisons between groups.

First research question: Does the educational level of English as an L2 affect the percentual usage of frequency levels of words used in writing?

Second research question: To what extent do lexical density and type-token ratio reveal differences between L2 writers from different educational levels?

4.1 Descriptive observations

In this section I will briefly go through how the over-all results look like without looking at specific education levels, but rather how all of the groups did together as a large group. Table 3 includes all the percentual usages of each K-band, NAWL and off-list, each calculated using word types.

Table 3. Statistics of how the groups together used different word types

	N	Mean	Median	Mode	Minimum	Maximum	Std. Deviation
K1 type %	147	81.60	83.20	85.4	51.7	95.5	7.84
K2 type %	147	6.69	6.50	8.80	0.0	16.8	3.35
K3 type %	147	2.37	1.90	0.00	0.0	8.0	1.89
NAWL type %	147	0.67	0.00	0.00	0.0	4.2	0.93
Off-list type %	147	8.07	7.00	3.4[a]	0.0	35.7	5.51

a. Multiple modes exist. The smallest value is shown

On average 81.6% (SD 7.84) of all types used across all texts belong to K1-band with notable results in all other categories as well. When looking at K2-band a drastic drop can be seen in the average compared to K1-band with 6.69% (SD 3.35). Also, when looking at the minimum

and maximum results one sees that minimum was zero and maximum is 16.8, which is over double the average results. These types of changes come even more apparent looking at the K3-band with an average of 2.37% (SD 1.89) and with again a minimum of zero and maximum this time over three times the average result with 8.0. Mode K3-band is also worth mentioning since it also is only zero. Moving on to NAWL types, here is an average of 0.67% (SD 0.93) and once again minimum of zero and maximum 4.2, which is over six times the average result. Mode again is zero but this time also the median is zero showing that a very slim portion of types used across all texts belonged to the NAWL group. Off-list types had an average of 8.07% (SD 5.51) with a minimum of zero and a maximum of 35.7, which again, is multiple times larger than the average.

Moving on to table 4, which includes the same data as table 3 but from the viewpoint of type-token and lexical density.

Table 4. Statistics of TTR and lexical density of all the groups together

	N	Mean	Median	Mode	Minimum	Maximum	Std. Deviation
TTR	147	0.59	0.60	0.61	0.41	0.72	0.07
Lexical density	147	0.50	0.50	0.52	0.36	0.67	0.05

In the case of TTR and lexical density the variation in the combined results of the groups were not all that significant, as can be seen from the standard deviation of 0.07 and 0.05, respectively. Almost no variation can be found between the mean, median and mode as well with only a 0.02 range of variation in each of the variables. When looking at the minimum and maximum, some variation can be seen with TTR's minimum being 0.41 and maximum 0.72, and lexical density's minimum being 0.36 and maximum 0.67.

4.2 Comparison between groups

In this section I am going to go through the results regarding the first research question: Does the educational level of English as an L2 affect the percentual usage of frequency levels of words used in writing? These results include the percentual averages of K1-, K2- and K3-bands counted using word types. Moreover, the percentual averages of NAWL-types and off-

list types are included in the thesis' results. Furthermore, I will also go through the results of Kruskal-Wallis test conducted to the different K-bands, as well as its post-hoc pair wise comparison results with the help of illustrative figures, which will be explained as they appear. I will go through different K-bands one by one starting with K1-band. I will pair the Kruskal-Wallis test and the post-hoc pair wise comparison results with their respective K-bands and go through them in their corresponding sections.

Table 5 shows the average percentages of all K-bands counted as word types as well as the average percentages for NAWL-types and off-list-types. These averages are divided further according to their respective education levels shown in the table's columns. Lastly, on the very right of this table are the p-values of each type.

Table 5. Averages of different education levels with their significances

	Comprehensive school	High school	University	Significance
K1 type % average	82.57	85.08	77.71	<0.001
K2 type % average	4.83	6.71	8.62	<0.001
K3 type % average	2.12	1.49	3.34	<0.001
NAWL type % average	0.67	0.39	0.89	0.005
Off-list type % average	9.16	5.89	8.73	0.007

Table 5 demonstrates all the mean ranks of the percentual usage of different K-types as well as NAWL-types and off-list-types with their corresponding education levels. Furthermore, there are also the p-values of the percentual usages. Next, I will go through the table 5 as well as the table 6 in more detail, paired with each K-band's respective post-hoc pair wise comparison results illustrated by triangle figures.

Table 6. Mean ranks and significances of different education levels

	Comprehensive school	High school	University	Significance
K1 type % mean rank	82.70	93.13	49.03	<0.001
K2 type % mean rank	49.88	75.40	93.38	<0.001
K3 type % mean rank	69.81	54.68	94.35	<0.001
NAWL type % mean rank	71.07	61.76	87.18	0.005
Off-list type % mean rank	77.31	57.11	84.41	0.007

Starting with the K1-types we can see that high schoolers used most K1-types in their answers with 85.08 % of their word types having been K1-types. Comprehensive school comes not that far behind with 82.57 % but university has a clear decrease of K1-type usage with 77.71 %. This is illustrated with a bar chart below in figure 1.

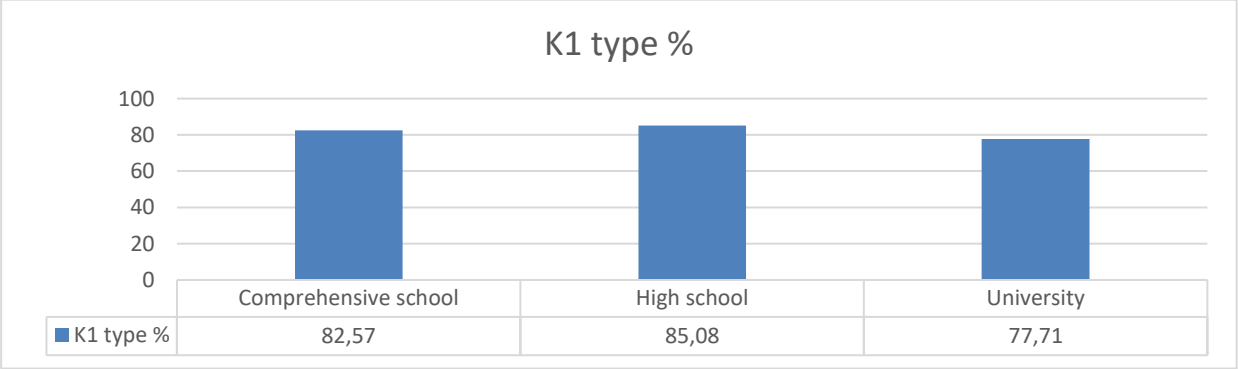


Figure 1. Percentual usage of K1-types across education levels

As for K1-type Kruskal-Wallis test and post-hoc pair wise comparison results, the Kruskal-Wallis test showed that there were statistically significant differences in the usage of K1-types between different levels of education, H (2) = 28.278, p = <0.001 with a mean rank usage of 82.70 for comprehensive school, 93.13 for high school and 49.03 for university.

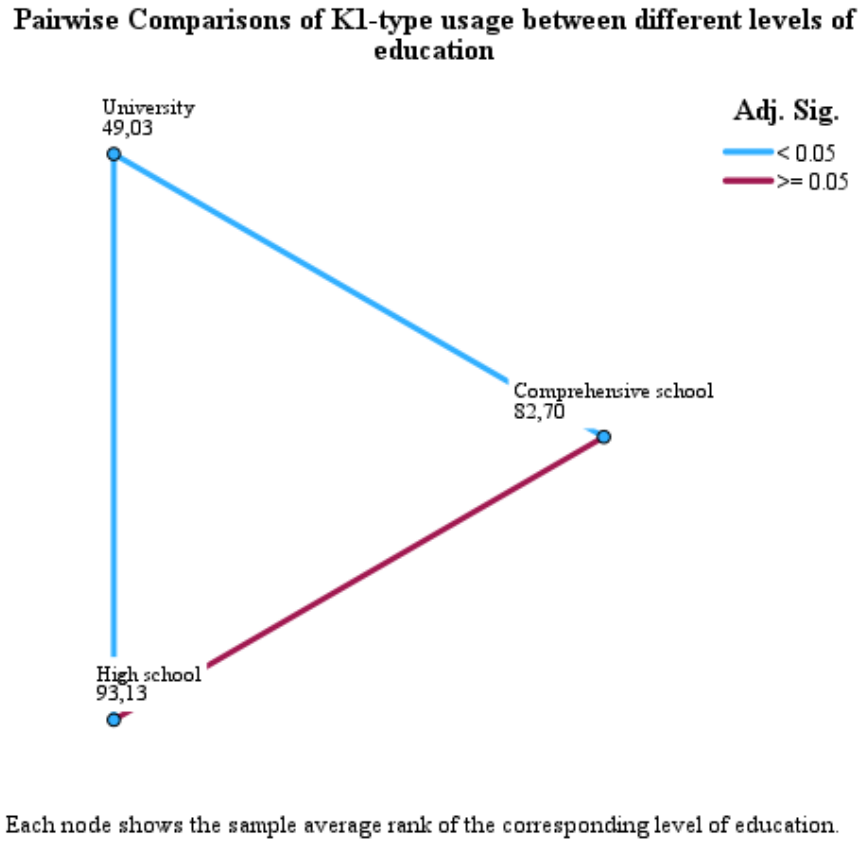


Figure 2. Pairwise comparison of percentual K1-type usage between education levels

Figure 2 consists of a triangle showing nodes with the mean ranks of all different levels of education and blue or red lines connecting these nodes. A blue line indicates that there was a significant difference between two connected groups while a red line indicates that there was not a significant difference.

Figure 2 more specifically shows the results of the conducted post-hoc tests to test pairwise comparison of the percentual usage of K1-types. As one can see from the blue lines of the triangle there were significant differences between university (77.71 %) and comprehensive (82.57 %) school as well as university (77.71 %) and high school (85.08 %) with adjusted p-values of 0.000 and 0.000, respectively. Likewise, as one can see from the red line of the triangle there was not a significant difference between comprehensive school (82.57 %) and high school (85.08 %) with an adjusted p-value of 0.702. Therefore, it can be said that university students outperformed both of the other two education levels. Even though

comprehensive school students used fewer K1-words than the high school students no significant difference was found. Thus, it cannot be said that one would have outperformed the other. This is because the fewer one uses K1-words the richer their language becomes i.e. the smaller number of K1-words in the text, the better the result. The percentages of the according word type usage are mentioned after each group in brackets for easier comparison, which will be a common feature through the results of this thesis.

Moving on to the percentual usage of K2-types figure 3 displays the percentages of this category.

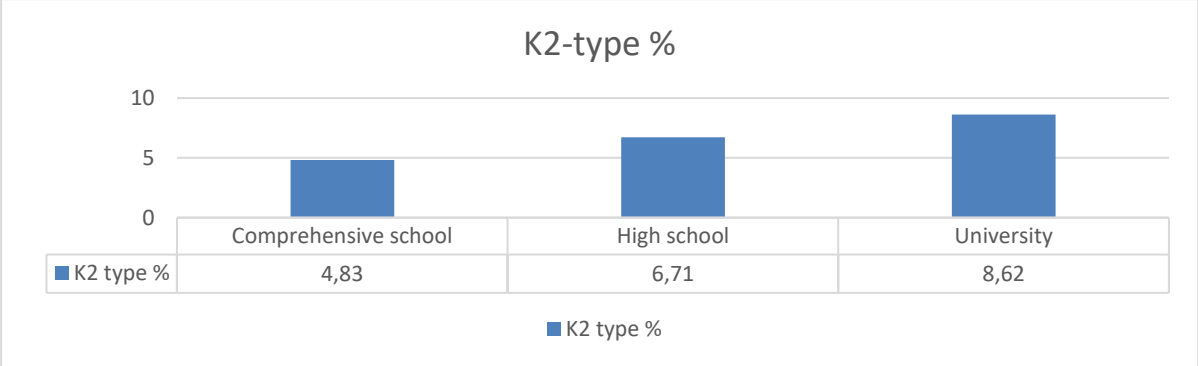


Figure 3. Percentual usage of K2-types across education levels

Here one can see that, unlike the K1-types, the results follow a linear model. In the low-end there is comprehensive school with 4.83 % of their used word types being K2-types, high school in the middle with 6.71 % and university in the high-end with 8.62 %. The Kruskal-Wallis test did show statistically significant differences between the education levels, H (2) = 32.106 and p = <0.001, with mean rank percentual usage of 49.88 for comprehensive school, 75.40 for high school and 98.38 for university.

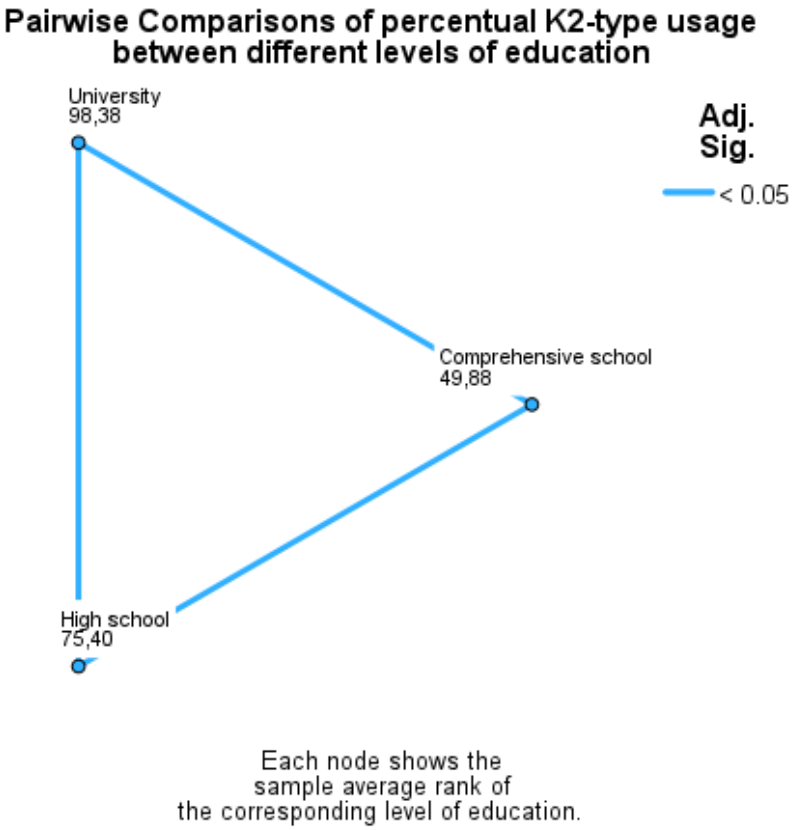


Figure 4. Pairwise comparison of percentual K2-type usage between education levels

Here an all-blue triangle is represented, meaning that between all level of education there was a significant difference with exact adjusted p-values of 0.011 between comprehensive school (4.83 %) and high school (6.71 %), 0.000 between comprehensive school (4.83 %) and university (8.62 %) and 0.029 between high school (6.71 %) and university (8.62 %). Thus, the results reveal that comprehensive school students used the least number of K2-types followed by high school students and university students used more than the two education levels. With K2-words the more are used the better the results, which is the other way round compared to K1-words.

Next is the last K-band, the K3-types, and its results are in a way an inverse version of K1-results.

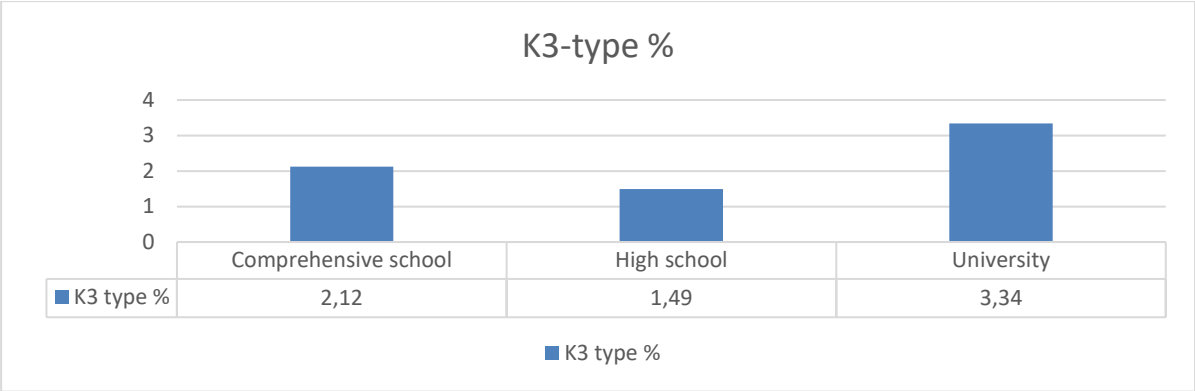


Figure 5. Percentual usage of K3-types across education levels

Figure 5 demonstrates K3-type usage and while in K1-type usage high schoolers had the highest usage of the three this time they had the lowest usage with 1.49 %. The same reverse logic applies to university students who had the lowest usage in K1-types, now they have the highest usage when looking at the K3-types with 3.34 %. The comprehensive school students can be found from the middle with K3-type usage of 2.12 %. The Kruskal Wallis H test showed that once again there was statistically significant difference in the percentual usage of K3-types, H (2) = 20.966, p = <0.001 with mean rank usage scores of 69.81 for comprehensive school 54.68 for high school and 94.35 for university. Looking at figure 6 depicting this K-band the following is apparent.

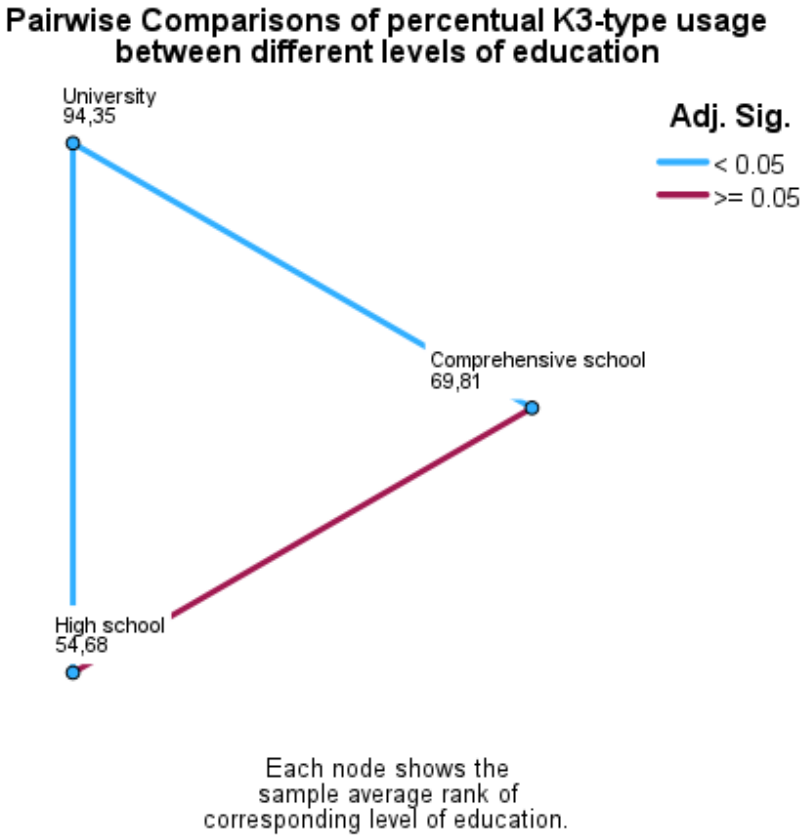


Figure 6. Pairwise comparison of percentual K3-type usage between education levels

There was a significant difference between university and the other two levels of education with adjusted p-values of 0.009 between comprehensive school (2.12 %) and university (3.34 %) and 0.000 between high school (1.49 %) and university (3.34 %). This time there was no significant difference between comprehensive school (2.12 %) and high school (1.49 %) with an adjusted p-value of 0.249. Therefore, it can be said that, like in K1-words, university students outperformed both other two education levels and that even though comprehensive school students used more K3-words the high school students, no significant difference was found. Thus, again, it cannot be said that one outperformed the other. This is because like with K2-words, the more K3-words are used the better the results are from the viewpoint of lexical richness.

Continuing to the percentual usage of NAWL-types, figure 7 shows that the results for the NAWL-types follow the same pattern as K3-types. High schoolers having the lowest percentage with 0.39 % of their word types belonging to the NAWL. University students had

the highest percentage with 0.89 % and comprehensive school students can again be found the middle of the three with 0.67 %.

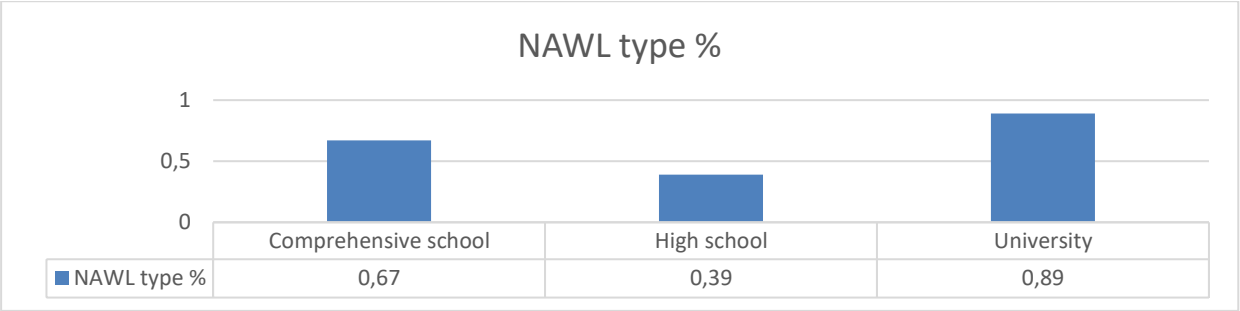


Figure 7. Percentual usage of NAWL-types across level of education

The Kruskal-Wallis test showed that statistically significant difference was found with H (2) = 10.767 and $p = 0.005$ with mean ranks of 71.07 for comprehensive school, 61.76 for high school and 87.18 for university.

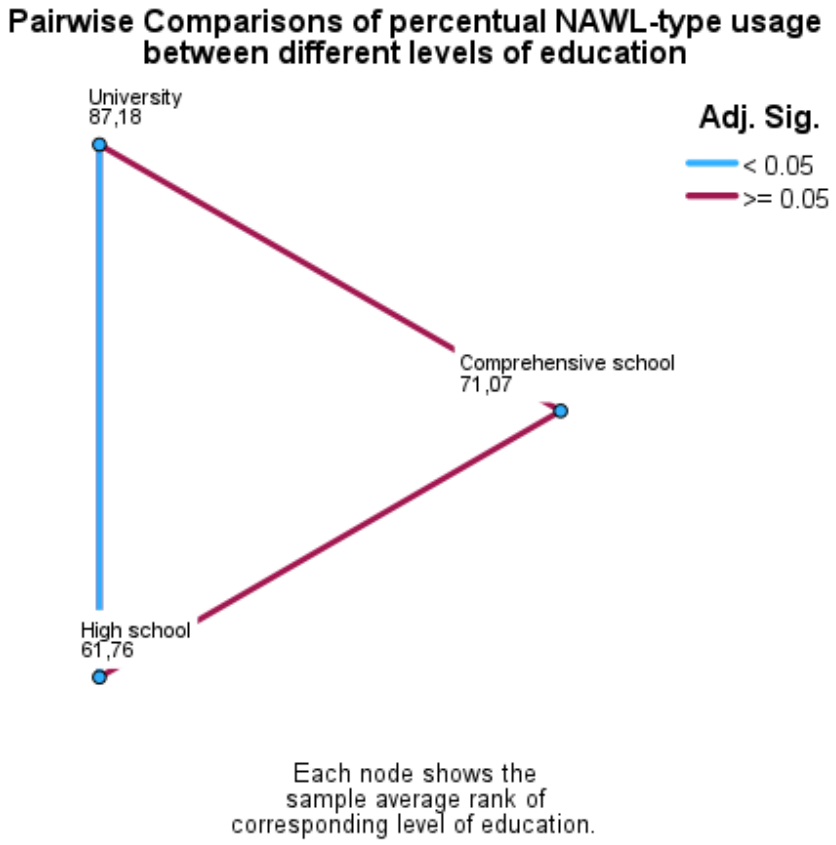


Figure 8. Pairwise comparison of percentual NAWL-type usage between education levels

Figure 8 illustrates the results of pairwise comparisons of how NAWL-types were used and this time there were two relations that had no significant difference and just one that had. The one that had was between university (0.89 %) and high school (0.39 %) with an adjusted p-value of 0.004. The ones that had significant differences were between comprehensive school (0.67 %) and university (0.89 %) and comprehensive school (0.67 %) and high school (0.39 %) with respective adjusted p-values of 0.703 and 0.091. These results reveal that even though, the university students used the most NAWL-words, they only outperformed comprehensive school students because no significant difference was found between them and high school students. Because of no significant difference between high school students and comprehensive school students it cannot be said who outperformed who. This is because the more NAWL-words are used the richer, or at least more academic, the text becomes.

Lastly there is the percentual usage of off-list-types. As figure 9 shows comprehensive school students had the highest percentage of off-list-types in their answers with 9.16 %. In the middle this time are university students with 8.73 % and the lowest off-list-type usage was found in high school students with 5.89 %.

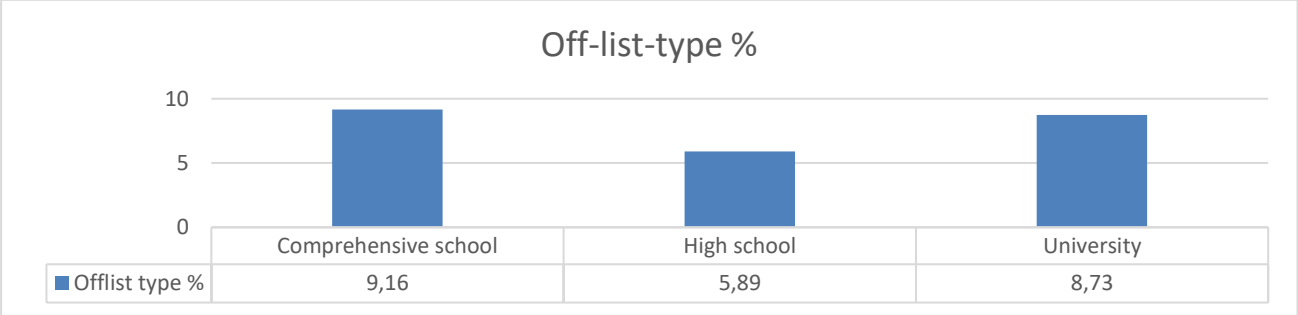


Figure 9. Percentual usage of off-list-types across levels of education

Just like in the others before this, Kruskal-Wallis test showed significant difference between the groups with H (2) = 9.988 and p = 0.007. Mean ranks were 77.31 for comprehensive school 57.11 for high school and 84.41 for university. Below is figure 10 showing the results of pairwise comparisons between the groups.

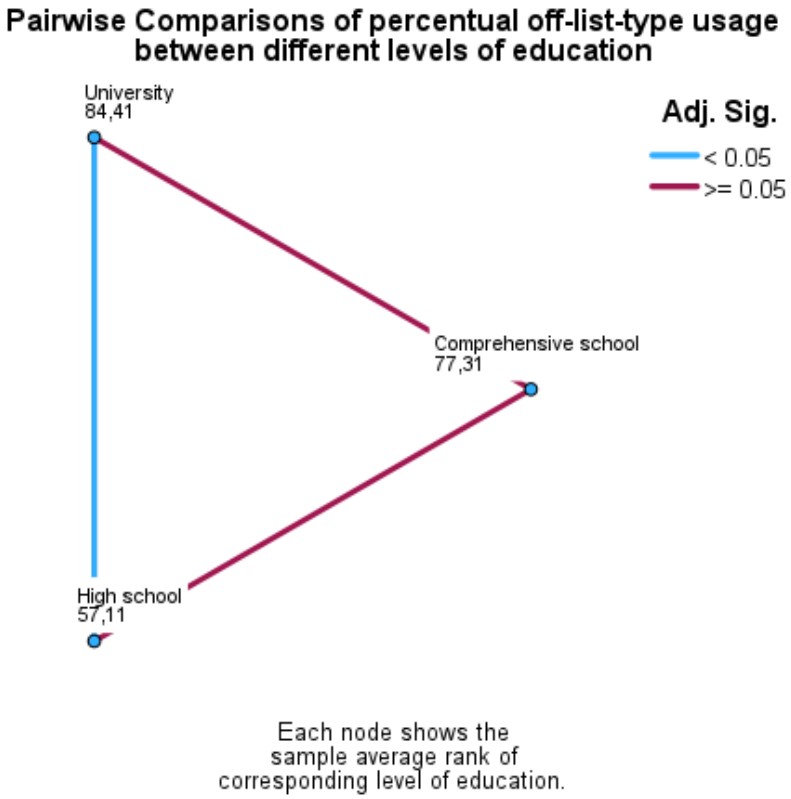


Figure 10. Pairwise comparison of percentual off-list-type usage between education levels

This triangle shows similar results to the last one with only one significant difference and that being between high school (5.89 %) and university (8.73 %) this time with an adjusted p-value of 0.006. Following this, like the last triangle, there are two relations that have no significance and are between comprehensive school (9.16 %) and high school (5.86 %) and comprehensive school (9.16 %) and university (8.73 %) with adjusted p-values of 0.061 and 1.000, respectively. The same kinds of conclusions cannot be made about the off-list-words as the previous ones, because it is not inherently clear what the off-list includes. This is a topic that will be discussed in more detail in the discussion.

4.3 Type-token ratio and lexical density

In this section I will give results that answer the second research question: To what extent do lexical density and type-token ratio reveal differences between L2 writers from different educational levels? These will include type-token ratios and lexical densities of each level of

education paired as well as Kruskal-Wallis test results paired with pairwise comparison results to show where the significant differences lie, if lie at all.

Starting with TTR as figure 11 shows, there was almost no variation between the groups. Comprehensive school students had the highest TTR with 0.60, followed by university students with 0.59 and high school students had the lowest with 0.56.

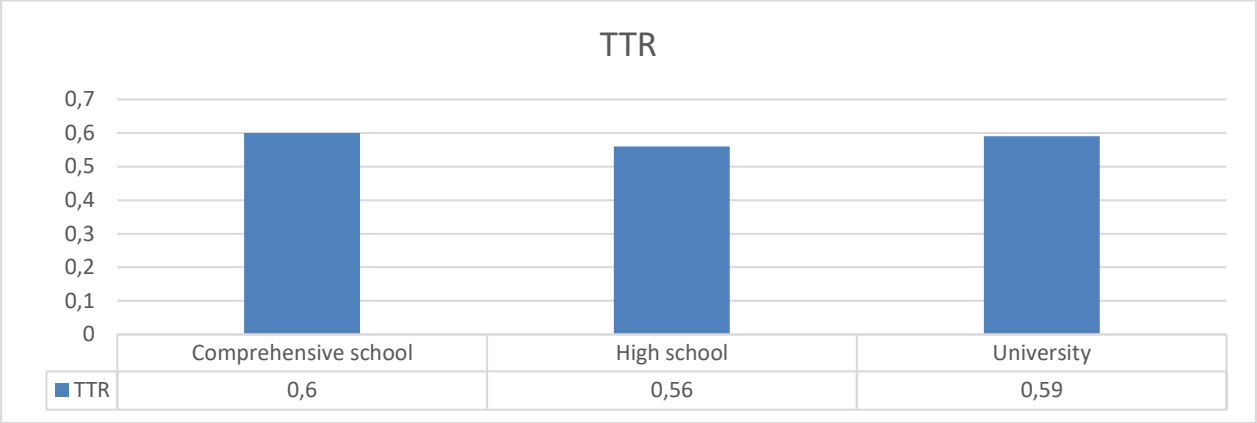


Figure 11. Type-token ratio results between different levels of education

The Kruskal-Wallis test did show significant difference between the groups with H (2) = 7.072 and $p = 0.029$. Mean ranks were 83.18 for comprehensive school, 60.12 for high school and 75,72 for university.

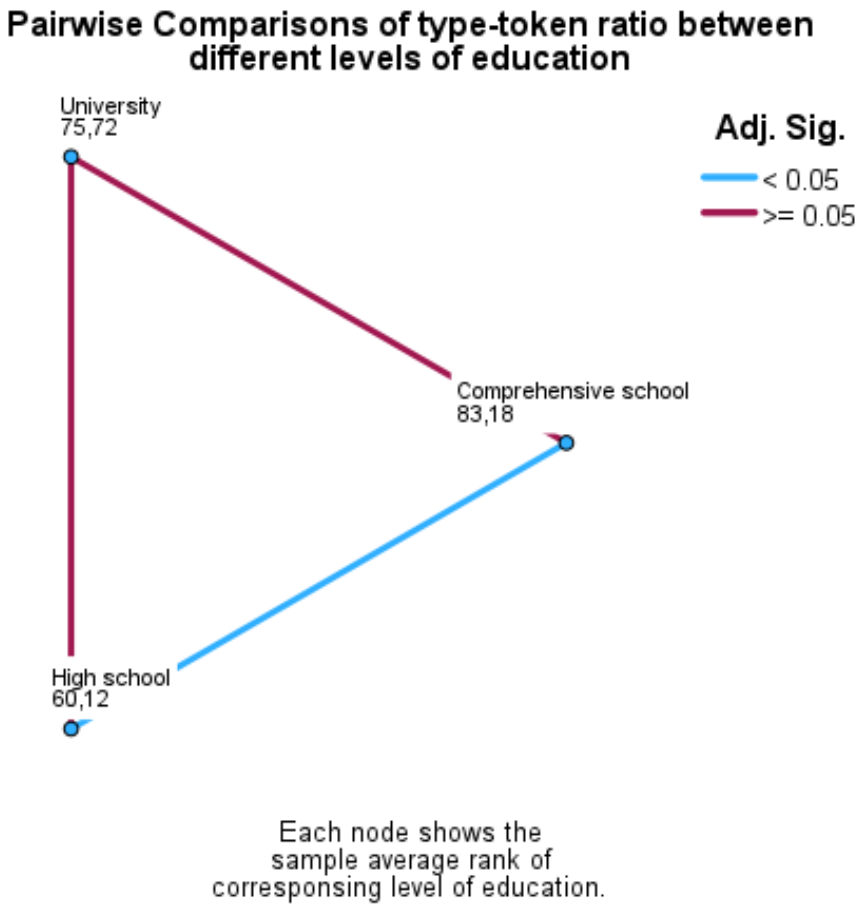


Figure 12. Pairwise comparison of type-token ratio between education levels

Figure 12 shows that only one pair had a significant difference between them, and it was between comprehensive school (0.60) and high school (0.56) with an adjusted p-value of 0.025. Between comprehensive school (0.60) and university (0.59) and high school (0.56) and university (0.59) there were no significant differences with adjusted p-values of 0.235 and 1.000, respectively. Therefore, it can be said that there are next to no differences between the groups when looking at type-token ratio.

Moving onto lexical density, figure 13 shows that there was even less variation between groups in lexical density than type-token ratio. High schoolers had the highest lexical density with 0.50. Comprehensive school students and university students had the same lexical density with 0.49, only a 0.01 decrease from the high schoolers.

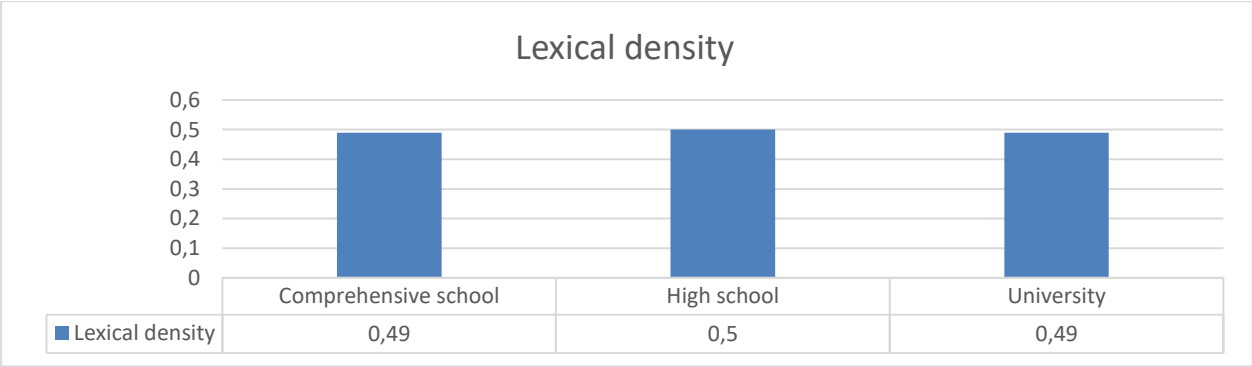


Figure 13. Lexical density results between different levels of education

The Kruskal-Wallis test showed no differences between the groups with H (2) = 1.243 and *p* = 0.537. Thus, it can be concluded that the performance across all of the education levels was extremely similar.

5 Discussion

As the results were discussed in the previous chapter in their numerical form without going too in depth to the findings of the present study, this chapter is dedicated for diving deeper into these findings and giving more light for why the results came to be. Just like the previous chapter, this chapter is divided into two main parts corresponding with the research questions of this study. Firstly, I will discuss the first research question and features and questions linked to it in light of previous studies. These will include, but not limit to, high schoolers performance, the off-list and its interpretation and vocabulary analysers and their short comings. Continuing from there I will move onto discussing the second research question and shed more light on to what extent lexical parameters such as type-token ratio and lexical density reveal differences between education levels, if at all.

5.1 K-bands and the NAWL

Starting with different K-bands and NAWL and how their usage changed between groups it can be seen that for some reason comprehensive school students performed better than high school students despite having studied English for lesser amount of time. This can be said because when Laufer and Nation (1995) showed that lexical frequency profiles can discriminate between different proficiency levels of English in their results the more proficient one's English knowledge was the less they used K1-words. The same applied to K2-words, not to the same degree but still some decrease was found. In Laufer and Nation's results academic words saw a significant increase as well as the off-list-words. Comparing those results to that of the present study the complete opposite can be seen when comparing comprehensive school students and high school students. In both cases high school students used less of the said categories in their compositions even though on the basis of Laufer and Nation's results the opposite could be assumed.

K3-band was skipped till this point because as Laufer and Nation used a different corpus, the General Service List, opposed to the New General Service List used in the present study. The General Service List did not include K3-band and as far as I could research, no significant studies of the nature of the present study and using the same corpora have been conducted at the time of writing. Because of this, only assumptions can be made how K3-band usage should develop as the time of studying is increased, but that does not take away from the results of this study. The results of off-list-words from Laufer and Nation's study does give a

standing point for my main argument regarding the K3-band. As discussed previously, the words that end up in the off-list are words that were not found in corpora used to calculate one's lexical frequency profile. From word frequency's point of view the only place the off-list words can go to are less frequent K-bands, excluding the academic words, which are not on the off-list. As discussed in chapter 2 and illustrated in table 2, we know that the coverage of K3-band (4.3 %) is significantly larger than coverage of K4-band (2.8 %) and that this percentage rapidly decreases as K5-band's coverage is only 1.9% and K6's 1.2 %. This means that just by chance it is more likely for the off-list-words to fall into K3-band rather than even less frequently used K-bands. As the percentual part of the off-list increased in Laufer and Nation's study as the proficiency level increased it could be assumed that a considerable portion of the off-list-words are K3-words. Thus, it could be assumed that increased usage of K3-words should be seen as the proficiency level of the participant increases.

If one was to exclude the university students results from the present study and looked at the results, complete opposite would be apparent. This is contradictory to the argument I just proposed but as discussed previously so are all of the results when comparing the results of the present study's comprehensive school students to high school students. Thus, the argument is coherent. Furthermore, present study's third group's results, the university student's results, give more evidence that the proposed argument regarding K3-band would be sound. This is because when comparing the results of the university students results to the other two groups, they follow the same pattern as Laufer and Nation's results did. The only two exceptions being that university students did use more K2-words than the other two groups even though, according to Laufer and Nation's study, they should have used less of them compared to the other two and the university students used less off-list-words when they should have used more of them. The off-list results of the present study will be revisited later in this chapter to give possible explanations for these results. To provide possible further evidence for the discussed claims there is a study by Paul Pauwel (2015), which used different corpora called BNC/COCA, which goes up to the K3-bands and works with the same principles as the corpora of the present study. Here the participants were 18 first year bachelor students of English who wrote three different compositions at three different times all of which were done in the course of six months. In this study an increase in K3-word usage was found with increasing proficiency level. In the first composition K3-word usage was 3.74 % compared to the last composition, which had a usage of 5.15 %. Contradictory to Laufer and Nation's study but in line of the results of the present study, this study also found an increase

in K2-word usage as the proficiency level improved. In the first composition 10.21 % of the words belonged to the K2-band and in the last 12.04 %.

Taking a closer look at the percentual usage of NAWL-words in the present study and comparing these results to the results of Laufer and Nation's study a clear difference can be seen. In Laufer and Nation's study all of the participants were university students, but the least proficient group were not English students but rather studied some other subjects and "were found to be at a low intermediate level". The next group was English students two were in their first year of English studies and the third group consisted of students who were in the end of their second semester. In their study the results, regarding academic word usage, were from least proficient group to most proficient, 3.2-4.1 %, 7.8-8.1 % and 8.1-10.1 %. Comparing these numbers to the results of the present study an enormous difference is clear as in the present study the same results in the corresponding order were 0.67 %, 0.39 % and 0.89 %. All groups in Laufer and Nation's study used over four times and up to 11 times more academic words than the university students in the present study. The main difference that cannot be ignored is that the studies used different corpora to measure academic word usage, which is a factor of which's effect on the results is hard to estimate but as the combined coverage of the NGSL and the NAWL is about five percent higher than the coverage of the GSL and the AWL it can be assumed that NAWL is at least as proficient of a tool at measuring academic word usage. The difference in coverage was tested in the 288-million-word academic corpus that the founders the NGSL and the NAWL used to develop said corpora (New General Service List Project 2023). Laufer (1994) found similar results to the results of Laufer and Nation where the data in question was entrance exam answers to study English at university level where the participants had to write a composition in English. Laufer found that 5.39 % of these answers consisted of academic words. The only somewhat reasonable explanation for the discrepancies between the results of the present study and the results of the two other studies could be found in the topics of the compositions.

Topics are an extremely important part of these kinds of studies because they automatically give the guidelines to what type of language is used in the raw data, from which the results are drawn from. In the case of Laufer and Nation (1994) the participants wrote two compositions. On the first one the participants were not given a choice for topics, and they all wrote on the same topic. In the second one they had three different choices and all of these choices, the first composition's topic included, were described to be "of a general nature and dealt with controversial issues". They also did not require expert knowledge of any particular subject

matter. In Laufer's (1994) study the compositions were only described as free written compositions for entrance exams. Pauwels (2015) had to rerun their study because they found the correlations between measures were not stable across essays. In this study the participants had three different choices in the first run. These were described as "changing clock in summer to winter; job-dependent retirement age; entrance exams for HE". In the case of students scoring above average either no progress or less spectacular progress was found when writing about clock change and below average writes had more progress when writing about the topic at hand. This made the researcher rerun the study and removing clock change as a topic and replacing it with "heavier punishment for driving offenses" as a topic after which the results were consistent.

It might be the case that the topics for the present study's compositions for some reason did not elicit academic language. In the case of comprehensive school students, it could be argued that their vocabulary just does not include that many words that belong to the NAWL but in the case of high schoolers, and especially in the case of university students, it is hard to believe that their vocabulary would include only a fraction of NAWL-vocabulary that of the participants of the studies discussed.

One of the biggest limitations of the present study is the off-list. As discussed multiple times, off-list-words are words that did not appear in the NGSL nor the NAWL. If the compositions were not prepared, most of the off-list-words would be misspellings. This was not the case in the present study, which means that all of the off-list-words should be words that belong to more infrequent K-bands, excluding the NAWL-words. An oversight with proper nouns happened when preparing the data. They were not removed as probably should have been done, which means that almost all of the proper nouns used in the compositions flooded the off-list and skewed the data in that regard.

This was especially apparent in compositions written by comprehensive school students and even more apparent when they chose to write about the person they admired. Comprehensive school students had the habit of explaining, sometimes in great detail, places the person has visited and people they have worked with. This kind of answering elicits a lot of proper nouns, which almost always fall under the off-list-category.

Example 1. Prepared version of a comprehensive school student's composition

> A person I is Teemu Selänne
> Teemu Selänne is a Finnish hockey player and he plays still at NHL. He is 42 years old and he has been voted to be the best Finnish hockey player. Teemu started his career in the year 1986 at Helsingin Jokerit. He was one of the best players at sm-liiga. Winnipeg Jets reserved him at 1988. When Teemu started at Winnipeg he scored 76 goals at his rookie season. Teemu has won four times the Maurice Richard trophy. That means that he has been the best goal scorer in the league four times. He plays nowadays at Anaheim Ducks in the NHL. He plays with his very good friend Saku Koivu. Teemu is the best Finnish guy on the score sheet with Valtteri Filppula they both have scored 23 goals and assisted 39 goals. I him because he is 42 years old and is one of the best players in NHL still.

Example 1. is a great example of this kind of answer. There was a total of 83 types used in the answer out of which 24 (31.2 %) fell under the off-list and out of these 24 types 18 were proper nouns, which comes out to a total of 21.7 %. Were the proper nouns removed, the compositions would have had 65 total types out of which 6 were off-list-types meaning that only 9.5 % of all the types would have been off-list-types. This is an extreme example and luckily most of the cases had only one to three more proper nouns compared to the compositions of university students and high school students.

To give perspective what the off-list should look like in a more ideal situation I present example 2, which again is an extreme example but demonstrates excellently why the off-list is a possible data point when differentiating between proficiency levels.

Example 2. Prepared version of a university student's composition

> A person I
>
> If admiring someone was merely a question of appearance, I would probably go for some picturesque movie star like the girls of my age are prone to. However, since I find it quite superficial, to put it mildly, to focus on the exterior only, I would rather opt for the legendary Oscar Wilde as the object of my unconditional worship. Anyone who has encountered the lad's picture cannot honestly claim I would exhibit gross shallowness...
>
> Why Wilde in particular? Having only confronted his production last summer, one could claim I was easily enchanted. Cheesy as it may sound, it was indeed love at first reading. After finishing 'The picture of Dorian Gray', a book which to all intents and purposes altered my life, I went to great lengths so as to obtain the complete works of Wilde. Getting further acquainted with his production I also grew still more infatuated. By the time I had swallowed the whole volume I declared I had found my Mr. right - leastways literary-wise.
>
> My sweetheart wrapped me around his finger with his shrewd remarks and facts about humanity and life in general. I have yet to meet any other writer who would to an equal measure have unraveled the secrets of existence to me. Add to this Wilde's witty style and taunting humor, and you can not but raise your cap to the man.

Example 2. is a composition with the topic “A person I admire” written by a university student. Here the total amount of types was 159 out of which 30 (18.9 %) were off-list-types. Out of these 30 only three were proper nouns and the rest were words that most likely only an advanced level English learner would know, let alone use correctly. Removing the three proper nouns, the more accurate off-list percent would be 17.3 %.

Another aspect that skews the off-list’s results also has to do with the topics but not the topics themselves but rather with the choosing of the topics. Unfortunately, the rate at which each group chose each topic was not constant in the slightest. The topic “Computers in my life” was chosen a total of 38 times out of which 25 were comprehensive school students, only five were high schoolers and only eight were university students. “A person I admire” was chosen a total of 53 times, 22 for comprehensive school, only 7 for high school and 24 for university. The same data for “16-year-olds should be allowed to drive a car in Finland!” would be only 6 for comprehensive school, 30 for high school and 18 university. Comprehensive school students clearly favoured computers and admiration as topics, which, as discussed earlier, had a definite effect at least on the off-list’s results. This bias could be explained by two reasons. First, it could be argued that computers and admiration is a topic that most likely is either familiar or easy to answer for this age-group. Second, in Finland driving’s license can be acquired in normal circumstances at the age of 18 and comprehensive school students, in this case ninth graders, are in general only 15-year-olds. These together might make the topic seem further away from them compared to computers and admiration, which might have led them to prefer to choose the latter topics.

This same bias is most likely also the explanation for the most obvious discrepancy in the results. Why did the high schoolers perform worse than the comprehensive school students despite having done two to three more years of English studying? Unfortunately, high schoolers were the most bias of all the groups when looking at the topic choosing. Whilst the other two groups had two topics that were overrepresented, in the case of high schoolers, only one topic was overrepresented and the other two clearly underrepresented. The total number of essays written by high schoolers was 42 and only five of them dealt with computers, seven with admiration and the rest of them, 30, dealt with driving’s licenses. This most likely is because in Finland you, in general, enter high school on the year you turn 16 and you graduate on the year you turn 19. This means that a topic about should or should not 16-year-olds be allowed to drive a car in Finland might have been too close to the conditions of the lives of the participants in question, leading them to choose the topic a lot more often than the other participants.

The topic at hand can easily be seen as a reason for why high schoolers performed worse on the off-list-variable because of the caveat explained earlier, the proper nouns and the topic regarding admiration. It can easily be assumed that when talking about a person one admires or the driving's license procedures in Finland, the first one is more likely to include more proper nouns and because they were not removed in the present study, they skew the results. The same can not as easily be justified for the other variables though. It might just be that for some reason vocabulary used to discuss driving's licenses does not naturally elicit the use of rarer words but rather is easiest explained by more common words. The participants did not have any incentive to write as well as they could. The lack of motivation to show one's skills together with the vocabulary used to easily describe their thoughts on driving's licenses might be the main reason why high schoolers performed as weakly as they did. There are most likely a lot of reasons for the performance of the group in question, but it is hard to deny that the bias in topic choosing would not be one of the major contributors.

Reasons similar to the ones discussed could be argued for the other variations in topic selection as well but the main take away from this is that topics are an extremely important variable of these types of studies and should be chosen by the researcher with great care and most preferably through some kind of testing.

5.2 Discussion for type-token ratio and lexical density

The results for the second research question of the present study were not as various compared to the first's results. Starting with the aspect with next to no variation, lexical density. The results of this measure were almost identical between the groups, and it seems like it was completely unaffected by proficiency level. This however is not too surprising as Pietilä (2015) came up with similar findings when they looked at MA thesis conclusions written by Finnish and Czech students of English and compared them to similar text written by native speakers of English. In their study they found that both of these groups had the same exact lexical density of 0.54, which is very close to the results of the present study. In the case of the present study the lexical density was a bit lower with the results varying between 0.49 and 0.50. This is not surprising though because of the lower proficiency level of most of the participants in the present study and in the case of university students, they probably had less motivation and time to craft a piece of text comparable to MA thesis conclusion.

Moving on to lexical variation, the only statistically significant relation between the groups of the present study was between comprehensive school students and high school students. This is interesting because there are previous studies suggesting that lexical variation should to an extent discriminate between proficiency groups like was the case of Pietilä's (2015) study where the same groups discussed in the previous paragraph were studied. The non-native writers had a lexical variation of 0.48 and the native writers 0.38, compared to the present study's 0.60, 0.56, and 0.59, from the least proficient group to the most. Furthermore, similar results were found by Danelund (2013) where they compared high school students. First-year students were compared to third-year students by analysing free compositions written by these two groups. In this study the less proficient group had a lexical variation of 0.47, while the more proficient group had 0.44. This is contradicting to what was just said because these numbers suggest that the less proficient group would have had less repetition in their compositions. This is because text length has been shown to affect lexical variation because the longer the text, the greater the chance to repeat words. In the case of Danelund the less proficient group had an average text length of 258 words, while the more proficient had an average of 357. To combat this, further calculations were applied and after further research, it was found that the third-year students did have better lexical variation.

Text length is a variable that should have been taken into account in the present study as well. As the average text lengths were for comprehensive school students 116, for high schoolers, 143, and for university students 206, it is something that most probably does skew the results of lexical variation. The same type of further calculations should have been used as was done in Danelund's study.

An interesting thing about the present study's lexical variation results is the fact that for some reason the lexical variation, i.e. the repetition of words is extremely low compared to the two previous studies discussed. The higher the number is, the less times words are repeated and if one compared the least proficient group of the present study and the most proficient groups of the two other studies, one would see that for some reason the least proficient group has a much lower repetition rate than the others with 0.60 compared to third year high schooler's 0.44 and native speaker's 0.48. The high result of comprehensive school students could be explained by just the text length being so short but, even if one switched university students of the present study to their place the argument would still hold with a score of 0.59 and an average text length of 204 words. For this I honestly have no explanation other than the

possible the topics, which could affect the repetition rate, but it is extremely hard to believe that it would do it in this scale.

6 Conclusion

Vocabulary is a widely studied subject in linguistics and as technology has developed, new tools for more efficient research have been produced. The present study used one type of these tools to measure lexical richness in written compositions. The tools in question were two different vocabulary analysers, one provided by EAP Foundation and the other by a website called lextutor.com, the latter's name being VP-Classic. The present study has one main aim, to measure lexical richness, but it was divided in two research questions. The first one aimed to find out how the educational level of English affect the usage of different frequency levels. The second one aimed to find out whether or not lexical density and lexical variation revel differences between educational levels and if yes, to what extent.

This was carried out by acquiring 147 essays written by three different educational levels, about 50 per level. These educational levels were all from the Finnish educational system and listed from the least proficient to the most, comprehensive school students (ninth graders), high school students and university students of English. The participants had three different topics to choose from when they wrote the compositions and they ranged from around 70 to 250 words in length. The compositions were prepared by fixing all spelling mistakes to eliminate the effect on the results because of the vocabulary analysers would mistake them for rare words. For the same reason, contractions were opened up and unnecessary items like Finnish words and emojis were removed.

The vocabulary analyser provided by EAP Foundation was used to analyse frequency bands, or K-bands, as they are called. As corpora for the K-bands, EAP Foundation used the New General Service List and the New Academic Word List. VP-Classic on the other hand was used to calculate the lexical density and lexical variation of the essays.

The results of the study indicated that educational level does affect the percentual usage of different K-bands as well as the NAWL-words. When comparing the least proficient group to the most proficient a clear increase in performance can be seen in almost all the variables. All of the variables were tested by using word types rather than tokens to eliminate the effect of repeated words. K1-types saw a clear change comparing comprehensive school students to university students as the number of K1-types used in a composition is supposed decrease as the proficiency level increases. Of comprehensive school student's compositions 82.57 % were K1-types and for university students the same variable was only 77.71 %. In the case of

K2-types the same results were 4.83 % and 8.62 %. It is not clear by previous studies should this number decrease or increase as the proficiency level increases but in the case of the present study it saw a clear increase. In K3-types an increase was also found from 2.12 % to 3.34 %. The same trend continues in the NAWL-types as the least proficient group's usage was 0.67 % and the most proficient's 0.89 %.

Lexical density and lexical variation are measures calculated by counting how many of a particular type of words are in a text and comparing the number to another type of word. In the case of lexical density, the types are content words and function words and in lexical variation they are types and tokens. In lexical density the amount of content words is counted and divided by the total number of tokens in the text. In lexical variation the total number of types is counted and again divided by the total number of tokens. The results of these measures were the following. Six different relations across these two variables were compared between the groups and only one of them was statistically significant. In the case of lexical density, the results for the groups from least proficient to most proficient were 0.49, 0.50 and 0.49. None of these results compared to the other was statistically significant. For lexical variation the results were 0.60, 0.56 and 0.59. The only relation between these results that was statistically significant was between comprehensive school and high school. Because of this, in the case of the present study, these measures did not seem to reveal differences between educational levels.

Regarding limitations, the present study had two main limitations. Both of them are part of the reason why high schoolers were not discussed when explaining word frequencies in this chapter. First main limitation is that, as an oversight, proper nouns were not removed from the raw data used in the study. This means that especially the results of the off-list-variable are unfortunately skewed because proper nouns were used in such large numbers that made up a considerable amount of the off-list, especially in case of the comprehensive school students. This is most likely the reason why comprehensive school students performed above even the university students in the variable at hand. This is also probably the reason why high schoolers performed so poorly. There is another reason amplifying the effect of the first one, which is selection bias when taking a closer look at the choosing of the topics. Each group had their own clear preferences but in case of comprehensive school and university, their selection favoured two topics and diminished the last one. This was not the case with high schoolers because they favoured just one topic above the two other ones to the points that over 71 % of their compositions discussed just one topic. This topic had to do with driving's

license procedures in Finland and arguably had the smallest chance of natural use of proper nouns. This does not however explain the overall weak performance of the high schoolers. Unfortunately, there are no clear reasons behind why they performed as they did but it is hard to dismiss the effects of topic selection bias because of its drastic nature in the present study.

Considering the limitations of the present study and that this type of vocabulary research is fairly new because of the technological limitations of the past, not to mention the possible technological advantages of the future, some suggestions for possible future studies are presented next. There are different types of vocabulary analysers available for general use and these should be taken advantage of. The website that provides VP-Classic has multiple different analysers using different corpora and different methods. Analysers such as them and all alike can be an excellent tool for vocabulary research and should be used to provide more insight into the world of vocabulary research. This applies especially to vocabulary depth research as it is the less studied subject between vocabulary width and depth but can give great amounts of information when applying for example Laufer and Nation's (1995) Lexical Frequency Profile. As an improvement suggestion for these analysers an option to automatically opt out of proper nouns is proposed. This could be implemented via improving the analysers by coding them to be able to identify them by capital letters. This would cut out a large part of the work of removing them and to limit their effect in the future studies.

Topic selection bias is something that should be addressed in the future studies as well. As the present study and Pauwel's (2015) study have shown, the topic selection is a crucial point of these type of studies and should be considered with great care. Either the topics have to be extremely well picked by the researchers, or choosing process should somehow be limited to force the participants to choose topics with an acceptable variation to minimize possible bias that could skew the data. Another aspect dealing with the topics are the topics themselves and the type of vocabulary that is naturally used to provide a sufficient answer to them. This should also be considered and most preferably somehow tested, before conducting the main data collection of studies of this nature, as this is most likely a major reason, why in the present study high schoolers performed clearly worse than comprehensive school students. Some topics naturally might use different types of vocabulary, some more common, some rarer, and this could have a notable effect on the results and the possibility of such effect should be noted. In order to fill in the gaps of these types of limitations and to shed more light to these areas of research, more research should be done in this field.

References

Ala-Akkala M. 2010. *The Development of English Vocabulary during Upper Secondary School*. MA Thesis, University of Turku.

Bauer L. & Nation P. 1993. *Word families*. International Journal of Lexicography 6: 253-279.

Browne C., Culligan B. & Phillips J. 2023 *The New General Service List*. Accessed 27.12.2023. http://www.newgeneralservicelist.org

Carter R. 1998. *Vocabulary. Applied Linguistics Perspective*. London: Routledge

Coxhead A. 2000. *A new academic word list*. TESOL Quarterly 34 (2): 213-238

Danelund L. 2013. *Exploring the level and development of Danish high school EFL learners' receptive and productive vocabulary knowledge: A cross-sectional study of grade 10 and grade 12 Danish EFL learners' lexical knowledge*. Unpublished MA Dissertation, University of Copenhagen.

Francis N. & Kučera H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin.

Gass S. & Selinker L. 1994. *Second Language Acquisition: An Introductory Course*. Amsterdam: John Benjamins.

Goulden R., Nation P., & Read J. 1990. *How Large Can a Vocabulary Be?* Applied Linguistics 11, no. 4: 341-363.

Greenbaum S. & Quirk R. 1990. *A Student's Grammar of the English Language.* Essex. Longman

Henriksen B. 2006. *Exploring the quality of lexical knowledge in the language learner's L1 and L2*. In Pietilä, Lintunen & Järvinen (eds): 13-36.

Henriksen B. & Danelund L. 2015. *Studies of Danish L2 learners' vocabulary knowledge and the lexical richness of their written production in English*. In Pietilä, Doró & Pipalová (eds): 29-55.

Jaatinen S. & Mankkinen T. 1993. *The size of vocabulary of university students of English*. In Sajavaara & Takala (eds): 101-126.

Laufer B. 1992. *Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability?* Journal of Research in Reading 15 (2): 95-103.

Laufer B. 1994. *The Lexical Profile of Second Language Writing: Does It Change Overtime?* Relc Journal 25: 21-33

Laufer B. & Nation P. 1999. *A vocabulary-size test of controlled productive ability.* Language testing 16 (1): 33-51.

Makkonen H. 2008. *Measuring the vocabulary knowledge of advanced and intermediate English students at university level.* Unpublished MA thesis. Department of English Translation Studies, University of Turku

Meara P. 1996. *The dimensions of lexical competence*. In Brown, Malmkjaer & Williams (eds): 35-53.

Meara P. [1992] 2010. *EFL Vocabulary Tests.* Swansea: Lognostics.

Melka F. 1997. *Towards a new approach to modelling vocabulary acquisition.* In Schmitt & McCarthy (eds): 109-139.

Morris L. & Cobb T. 2004. *Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees*. System (Linköping), 32(1), 75–87. https://doi.org/10.1016/j.system.2003.05.001

Nation P. 2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nation P. & Meara P. 2002. *Vocabulary*. In N. Schmitt (ed.) An Introduction to Applied Linguistics Edward Arnold: 35-54.

Nation P. & Beglar J. 2007. *A vocabulary size test*. The Language Teacher 31 (7): 9-13

Nation P. & Hwang K. 1995. *Where would general service vocabulary stop and special purposes vocabulary begin?* System 23 (1): 35-41

Nation P. & Waring R. 1997. *Vocabulary size, text coverage and word lists*. In Schmitt & McCarthy (eds): 6-19

Pauwels P. 2015. *Lexical frequency profiles, lexical development and writing quality*. Accessed 7.3.2024 https://www.researchgate.net/publication/282322653_Lexical_frequency_profiles_lexical_development_and_writing_quality

Qian D. 1999. *Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension*. The Canadian Modern Language Review 56 (2): 282-307.

Read J. 2000. *Assessing Vocabulary* Cambridge: Cambridge University Press

Saville-Troike M. 2006. *Introducing Second Language Acquisition.* Cambridge: Cambridge University Press.

Schmitt N. 2010. *Researching Vocabulary: A Vocabulary Research Manual.* Hampshire: Palgrave Macmillan

Singleton, D. 1999. *Exploring the Second Language Mental Lexicon* (Cambridge Applied Linguistics). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524636

Stoeckel T., Ishii T. & Bennett P. 2020. *Is the Lemma More Appropriate than the Flemma as a Word Counting Unit?*, Applied Linguistics, Volume 41, Issue 4: 601–606, https://doi-org.ezproxy.utu.fi/10.1093/applin/amy059

Takala S. 1989. *Sanaston opettaminen ja oppiminen* (Vocabulary instruction and learning). Jyväskylä: University of Jyväskylä

Xue G. & Nation P. 1984. *A university word list*. Language Learning and Communication 3 (2): 215-229

Appendix 1 Finnish summary

Suomenkielinen tiivistelmä

Sanasto ja sen osaaminen ovat äärimmäisen keskeisiä asioita vieraan kielen oppimisessa. Tässä tutkimuksessa haluttiin keskittyä sanaston syvyyteen (vocabulary depth) sanaston laajuuden (vocabulary breadth) sijaan, joka on näistä kahdesta enemmän tutkittu sanaston osa-alue. Tutkimuksella on kaksi tavoitetta:

1) Kuinka kielellinen kompetenssi vaikuttaa sanojen frekvenssitasojen sekä akateemisen sanaston prosentuaaliseen käyttöön?
2) Missä mittakaavassa sanastollinen tiheys (lexical density) sekä sanastollinen vaihtelevuus (lexical variation) paljastavat eroja eri kompetenssitasojen välillä vieraan kielen oppijoissa?

Tutkimuksen teoreettinen tausta

Termi *sana* voidaan määritellä monin eri tavoin ja sen määrittely tilannekohtaisesti on tärkeää. Tämän tutkimuksen kannalta *sanan* määritelmät *flemma*, *tyyppi* sekä *token* ovat keskeisiä. Flemma sen takia, että tutkimuksessa käytetyt analysaattorien käyttämät korpukset määrittelevät sanan lemmaksi. Flemma on välimuoto *lemman* ja sanaperheen (word family) välillä. Lemmat lukevat samaan sanaan sen kaikki taivutusmuodot tarkoittaen sitä, että sanaa *lukea* voidaan taivuttaa esimerkiksi muotoihin, *luet, luin* ja *lukisimme* ilman, että yksikään näistä olisi eri lemma. Sanaperheeseen taas luetaan myös sanoista johdetut muodot eli käyttäen edelleen sanaa *lukea*, siitä voidaan johtaa edellä mainittujen sanojen lisäksi esimerkiksi sana *lukija* ja *lukematon*. Flemma on näiden kahden väliltä, koska lemmat erottelevat sanojen välillä, jotka kuuluvat kahteen eri sanaluokkaan, kuten numeraali *kuusi* ja substantiivi *kuusi*, mutta flemmojen kannalta katsottuna nämä kaksi ovat sama sana.

Token viittaa mihin tahansa jonoon kirjaimia, jotka on erotettu toisesta jonosta kirjaimia joko välilyönnein tai välimerkein. Tyypin määritelmään pätee samat säännöt kuin tokeniin, mutta tyypin kohdalla, jos sama jono kirjaimia esiintyy uudelleen annetussa tekstissä, sitä ei laskettaisi uudeksi sanaksi. Tyyppeihin lasketaan siis vain uniikit sanajonot. Nämä termit ovat tärkeitä tämän tutkimuksen kannalta, koska mahdollistaakseen mahdollisimman todenmukaisen käsityksen oppijoiden sanojen käytöstä on parempi laskea tyyppejä kuin tokeneita. Tämä siksi, että tekstiä, joka on 100 tokenia pitkä, mutta sisältää vain yhden tyypin, olisi vaikea pitää monipuolisena sanojen käyttönä. Toinen syy näiden termien tärkeydelle

tässä tutkimuksessa on sanastollisen vaihtelevuuden laskukaava, joka käyttää näitä molempia termejä.

Termin *sanasto* määrittelyyn lukeutuu myös monia käsitteitä. Yksinkertaisimmillaan *sanastolla* tarkoitetaan kielen komponenttia, johon kuuluu sanat ja niiden merkitykset. Kielitiede on nykypäivänä kuitenkin laajentanut ja syventänyt *sanaston* määritelmää ja nykyään sen ajatellaan olevan monipuolinen käsite, johon kuuluu paljon muutakin kuin vain sana-merkitys-pareja.

Kielitieteessä sanaston osaamisessa erotellaan tavallisimmin aikaisemmin mainitut sanaston laajuus sekä sanaston syvyys. Sanaston laajuudella viitataan yleensä sanaston kokoon eli siihen, kuinka monta sanaa kyseessä olevaan sanastoon kuuluu. Sanaston syvyydellä taas tarkoitetaan sitä, kuinka hyvin oppija hallitsee koko sanastoaan eli voidaan puhua sanaston laadusta. Sanaston laajuuteen taas kuuluu termi sanastollinen rikkaus, jota voidaan katsoa kolmelta eri kantilta: sanastollinen vaihtelevuus, sanastollinen tiheys sekä sanastollinen sivistyneisyys (lexical sophistication). Tässä tutkimuksessa keskitytään kahteen ensimmäiseen. Sanastollista vaihtelevuutta mitataan laskemalla tekstin tokenit ja tyypit ja jakamalla tyyppien lukumäärä tokenien lukumäärällä. Luku, joka tästä saadaan, on tekstin sanastollinen vaihtelevuus. Sanastollista tiheyttä mitataan laskemalla tokenit ja sisältösanat ja jakamalla sisällyssanojen lukumäärä tokenien lukumäärällä.

Sanastoa voidaan myös luokitella monella eri tapaa ja erityisen tärkeitä tämän tutkimuksen kannalta ovat frekvenssitasot sekä akateeminen sanasto. Frekvenssitasot ovat tasoja, joille sanat voidaan laittaa niiden yleisyyden ja harvinaisuuden perusteella. Ensimmäinen frekvenssitaso sisältää ensimmäiset 1000 yleisintä sanaa, seuraava taso taas 1001-2000 yleisintä sanaa ja kolmas taso 2001-3000 yleisintä sanaa, jatkuen teoriassa loputtomiin. Sanojen yleisyys ja harvinaisuus riippuu kieliaineistosta, josta sanojen yleisyyttä ja harvinaisuutta lasketaan. Tässä tutkimuksessa on käytetty kieliaineistoa nimeltä ”the New General Service List” (NGSL), joka on kehitetty englannin kielen opettelua varten ja sen tarkoituksena on kerätä kaikkein tärkeimpiä sanoja englantia vieraana kielenä opettelua varten (the New General Service List Project 2023). Se pohjautuu 273 miljoonan sanan kieliaineistoon, josta on johdettu 2809 tärkeintä sanaa englannin kielen opetteluun vieraana kielenä, jotka voidaan jakaa kolmeen eri frekvenssitasoon (ibid.). Akateemisella sanastolla tässä tutkimuksessa viitataan kieliaineistoon nimeltä ”the New Academic Word List” (NAWL), joka on aineisto kehitetty mahdollistamaan mahdollisimman tehokasta englannin

kielen opiskelua akateemisiin tarkoituksiin (the New General Service List Project 2023). Se on kehitetty toimimaan yhdessä NGSL:n kanssa ja ne yhdessä kattavat noin 92 % niiden kehityksessä käytettyä 288 miljoonan sanan akateemista kieliaineistoa. NAWL:ään kuuluu 957 sanaa ja se luetaan omana kategorianaan frekvenssitasojen ohessa.

Metodit

Aineisto, jota käytettiin tämän tutkimuksen toteuttamiseen, on osa Turun yliopiston vuonna 2014 keräämää suurempaa aineistoa, jota on tarkoitus käyttää kielentutkimukseen. Aineisto sisältää 147 esseetä, jotka ovat kerätty kolmelta eri koulutusasteelta, yläasteelta, lukiosta sekä yliopiston englannin kielen opiskelijoilta. Kaikki esseiden kirjoittajat ilmoittivat äidinkielekseen suomen kielen. 54 esseetä tuli yläastelaisilta, 42 lukiolaisilta ja 51 yliopisto-opiskelijoilta. Syitä vaihtelulle on muutama. Kaikilta asteilta ei alun perinkään ollut yhtä montaa esseetä saatavilla. Jokaisen esseen kirjoittajan piti antaa suostumus tutkimuskäyttöön ja esseiden minimipituudeksi asetettiin 50 sanaa.

Esseillä oli kolme eri aihevaihtoehtoa: ”Tietokoneet elämässäni”, ”16-vuotiaiden pitäisi saada ajaa autoa Suomessa!” sekä ”Ihminen jota ihailen”. Osallistujat saivat valita näistä aiheista vapaasti. Yläastelaisilla essee kirjoitettiin osana normaalia 45 minuutin englannin kielen oppituntia ja he saivat myös opettajan pyynnöstä avainsanoja sekä keskeisiä kysymyksiä, jotta tekstin tuottaminen olisi helpompaa. Heillä tekstin pituudeksi suositeltiin 80–100 sanaa. Lukiolaisilla ja yliopisto-opiskelijoilla suositeltu pituus oli 150–250 sanaa ja heille annettiin kirjoittamiseen aikaa myös 45 minuuttia.

Tutkimuksen aineistoa analysoitiin käyttämällä kahta eri sanaston analysointiohjelmaa. Verkkosivusto nimeltä EAP Foundation tarjoaa näistä ensimmäisen. Tällä analysointiohjelmalla ei ole erityistä nimeä, joten siihen viitataan vain kirjaimilla ”EAP”. EAP:tä käytettiin tutkimuksessa analysoimaan frekvenssitasoja. Ohjelman tulokset sisältävät viisi eri luokkaa: kolme eri frekvenssitasoa, NAWL-luokan sekä luokan, johon kaikki sanat, jotka eivät mene edellä mainittuihin luokkiin, niin sanottu ”off-list”. Toisena analysointiohjelmana käytettiin ohjelmaa nimeltä VP-Classic. Tämä ohjelma, kuten EAP:kin, on vapaassa käytössä internetissä ja löytyy sivustolta lextutor.com. Tätä ohjelmaa käytettiin, koska se automaattisesti laskee sanastollisen vaihtelevuuden sekä sanastollisen tiheyden, ominaisuus, jota EAP ei tarjoa.

Jotta tutkimuksen aineisto toimisi näiden analysointiohjelmien kanssa kuten pitääkin sitä piti käsitellä. Kolmea ohjetta noudatettiin aineiston käsittelyssä, jotka olivat seuraavat: 1. Supistetut muodot tulee avata. 2. Kirjoitusvirheet tulee korjata. 3. ylimääräiset itemit, kuten hymiöt ja suomalaiset sanat tulee poistaa. Näitä kolmea ohjetta sovellettiin jokaiseen esseeseen. Ohjeet oli pakko luoda, koska muuten tulokset olisivat olleet erittäin vääristyneet erityisesti off-list-luokan osalta, koska kaikki supistetut muodot, kirjoitusvirheet sekä ylimääräiset itemit olisivat automaattisesti menneet off-list-luokkaan, jonka seurauksena siitä olisi muodostunut epärealistisen suuri.

Tulokset ja implikaatiot

Tutkimuksen tulokset ensimmäisen tutkimuskysymyksen kannalta ovat selvät. Kaikkien ryhmien välillä on kaikissa muuttujissa tilastollisesti huomattavia eroja. Yläastelaiset käyttivät ensimmäisen frekvenssitason sanoja 82,57 %, lukiolaiset 85,08 % ja yliopisto-opiskeijat 77,71 (p = <0,001). Toisen frekvenssitason sanoja käytettiin samassa järjestyksessä 4,83 %, 6,71 % sekä 8,62 % (p = <0,001). Kolmannen frekvenssitason kohdalla tulokset ovat seuraavat, 2,12 %, 1,49 % ja 3,34 % (p = <0,001). Akateemisten sanoja käytettiin seuraavasti: 0,67 %, 0,39 % ja 0,89 % (p = 0,005). Ja vielä lopuksi off-list-luokan sanojen kohdalla samat tulokset olivat, 9,16 %, 5,89 % ja 8,73 % (p = 0,007).

Kun tuloksia tarkasteltiin tarkemmin, huomattiin, että kaikissa muuttujissa ei kaikkien ryhmien välisissä suhteissa ollut tilastollisesti merkittävää eroa. Ensimmäisen frekvenssitason kohdalla yläastelaisten ja lukiolaisten välillä ei ollut merkittävää eroa, kun taas kahden muun suhteen välillä oli. Toisessa frekvenssitasossa kaikkien suhteiden välillä oli tilastollinen ero, mutta kolmannessa tasossa tilanne oli sama kuin ensimmäisessä. NAWL-luokassa vain yliopiston ja lukion välillä oli tilastollinen merkittävyys, kun taas yliopiston ja yläasteen ja lukion ja yläasteen välillä tällaista ei löytynyt. Sama toistui off-list-luokassa.

Näiden tulosten perusteella voitaisiin sanoa, että lukiolaiset suoriutuivat näistä kolmesta ryhmästä kaikkein heikoiten, vaikka heillä on 2–3 vuotta enemmän englannin kielen opiskelua takana. Tälle löytyy mahdollinen selitys esseiden aiheista sekä niiden valitsemisesta. Esseiden aiheet ovat tämän tyyppisissä tutkimuksissa äärimmäisen kriittinen osa tutkimuksen toteuttamista, koska ne antavat rungon sille minkälaisia sanoja minkäkin aiheen käsittelyssä tulee todennäköisesti käytettyä. Lukiolaisten aiheen valinta oli äärimmäisen yksipuolista. Heidän 42:sta esseestään viisi käsitteli tietokoneita, seitsemän

ihailtua henkilöä ja loput 30 käsittelivät soveliasta ajokortin saamisikää. Tämä todennäköisesti rajoitti lukiolaisten sanaston käyttöä, koska voi olla, että ajokortista ja autoista puhuminen ei luonnollisesti tuota harvinaisempaa tai akateemisempaa englannin kielistä sanastoa.

Yläastelaisia ja yliopisto-opiskelijoita verratessa erot ovat kuitenkin selkeät. Kaikkien muuttujien kohdalla yliopisto-opiskelijat suoriutuivat yläastelaisia paremmin, lukuun ottamatta yhtä. Off-list-luokassa yläastelaiset suoriutuivat kaikkein parhaiten. Tälle selitys löytyy todennäköisesti taas esseiden aiheista ja niiden valitsemisesta. Yläastelaiset suosivat esseiden valinnassa tietokoneita sekä ihailtua henkilöä. Heidän 54:stä esseestään 25 käsittelivät tietokoneita, 24 ihailtua henkilöä ja vain kuusi käsitteli ajokorttia. Off-list-luokan kohdalla erityisesti kiinnostaa ihaillut henkilöt, koska, kuten on helposti ymmärrettävää, henkilöistä puhuminen tuottaa paljon erisnimiä ja niihin liittyen aineiston käsittelyssä tapahtui epähuomio. Niitä ei poistettu alkuperäisestä aineistosta kuten olisi todennäköisesti pitänyt. Sanaston analysointiohjelmia ei ole ohjelmoitu tunnistamaan erisnimiä, joten ne lähes poikkeuksetta päätyvät aina off-list-luokkaan. Tämän takia yläastelaisten off-list-luokasta merkittävä osa on erisnimiä, joiden käyttöä ei kuitenkaan pitäisi lukea sanastollisen rikkauden piiriin.

Toisen tutkimuskysymyksen kohdalla tulokset olivat myös erittäin selkeät, mutta päinvastaiseen suuntaan kuin ensimmäisen kohdalla. Sanastollisen tiheyden kohdalla tulokset aloittaen yläastelaisista siirtyen enemmän opetusta saaneisiin ovat 0,49, 0,50 ja 0,49 (p = 0.537). Tästä huomataan, että tulokset ryhmien välillä eivät olleet tilastollisesti merkittäviä ja olivat myös lähes identtiset. Tämä ei ole erityisen yllättävää, koska on aikaisempia tutkimuksia (Pietilä 2015), joiden tulokset ovat päätyneet myös lopputulokseen, että sanastollinen tiheys ei välttämättä eroa kompetenssitasojen välillä.

Sanastollisen vaihtelevuuden tulokset olivat seuraavat: yläaste 0,60, lukio 0,56 ja yliopisto 0,59 (p = 0.029). Ryhmien välillä oli siis tilastollisia merkittävyyksiä, mutta kun ryhmien välisiä eroja tarkasteltiin tarkemmin, huomattiin, että ainoa merkittävyys oli yläastelaisten ja lukiolaisten välillä. Tässä muuttujassa mitä isompi tulos, sen parempi, joten jälleen kerran yläastelaiset suoriutuivat periaatteessa kaikkein parhaiten näistä ryhmistä, vaikkakaan heidän ja yliopisto-opiskelijoiden välillä ei ollut merkittävää eroa. Nämä tulokset ovat tosin selitettävissä sillä, että tekstin pituus vaikuttaa merkittävästi sanastolliseen vaihtelevuuteen, sillä mitä pidempi teksti, sitä todennäköisemmäksi sanojen toistuminen tulee. Yläastelaisilla

oli ehdottomasti lyhyimmät tekstit keskipituudella 116 tokenia, lukiolaisilla keskipituus oli 143 tokenia ja yliopisto-opiskelijoilla 204 tokenia.

Tutkimuksen tulokset viittaavat siihen, että sanaston analysointiohjelmia voi käyttää tehokkaaseen sanaston tutkimiseen, mutta pitää olla tarkka, miten aineistoa käsittelee, jotta tulokset olisivat mahdollisimman realistiset. Tulokset viittaavat myös siihen, että tehdessään tällaista tutkimusta, erityistä huomiota kannattaa kiinnittää aiheiden valintaan ja siihen, miten tutkimukseen osallistujien antaa valita aiheita. Lisää saman tyyppistä tutkimusta tarvitaan, sillä teknologia on vasta aluillaan siinä, miten kaikin tavoin se voi avustaa kielitieteen tutkimusta.