# Lexical Richness in the Texts of Norwegian Lower Secondary School Students: The Effect of Specific Extramural English Activities.

MARKUS JØRGENSBORG

SUPERVISOR
Lenka Garshol

## Acknowledgements

# Abstract

This thesis explores the vocabulary of a selection of 8th grade students and the effect that Extramural English activities can have on their vocabulary knowledge. The vocabulary of the students was based on measurements of lexical richness, these were, lexical diversity, lexical density, and lexical sophistication. Other variables such as off-list words and total words were also looked at. The primary focus is on activities such as watching movies/tv-series, listening to audiobooks/radio programs/podcasts. However, other Extramural English activities are also examined with regards to their effect on vocabulary. In this study the data is collected through the TRAWL project, access to student texts, and their answers to a questionnaire about Extramural English activities have been granted through this project. The participants of this particular study consist of 80 Norwegian EFL learners, 36 male and 44 female. All of the participants are 8th graders from the Agder area, and it can be assumed they are near exclusively age 12/13. To analyze the texts of the students a program Cobb's (2023) Vocabprofile VP Compleat v2.6 was used. A statistical analysis was run on the relationship between total time spent on Extramural English and the lexical richness scores. The statistical analysis required groupings. Two methods of grouping the participants were used. When looking at the time spent on activities individually the time the participants spent on these activities were set as groups. Meaning there was five groups, group 1= 0 hours, group 2 = 0.5 hours, and so on. When looking at several of the Extramural activities combined or when looking at all of the activities combined, the students were divided into three groups, low, moderate, and high. The conclusion from the analysis was that there was no significant correlation between the time spent on activities and the lexical richness scores of the students.

# Contents

# List of Figures and Tables

# List of abbreviations

EFL – English as a foreign language

FTR – Flemma-token ratio

LD – Lexical density

LS – Lexical sophistication

L1 – First language

L2 – Second language

NAWL – New Academic Word List

NGSL – New General Service List

TTR – Type-token ratio

# 1.0 Introduction

In this master thesis, the TRAWL project is used to research the connection between the time 8th grade Norwegian students spend on certain extramural activities and the vocabulary level in their written texts. For the purposes of this investigation, their vocabulary is ranked with different values related to their texts' lexical richness. The activities that are analyzed are mainly on time spent: watching tv-series/movies with or without English subtitles and listening to audiobooks/radio programs/podcasts. However, more activities are later included, such as time spent: gaming, reading online, reading offline, chatting/writing/sms/, talking to someone in English, and other specified uses of English. These are discussed, and the activities are looked at collectively as time spent on Extramural English.

## 1.1 Background

The Norwegian society, job market, and culture are increasingly influenced by rapid globalization. Because of this, the English language plays a vital role in the necessary skills the students should be trained in. One of the most important aspects of any language is to be able to learn vocabulary. Because of multimedia technology advances, English is easily accessible and there are now more ways than ever for students to acquire language. Therefore, Extramural English is increasingly becoming a part of the language learning picture. This is where pupils come in contact with English outside of the English classroom (Sundqvist & Sylvén, 2016, p. 6). Students in Norway start watching English tv-series and movies from an early age; some even listen to podcasts and audiobooks. Also, movies and tv-series are generally not dubbed in Norway. These types of media allow for a lot of authentic English input, and they are also something that students use for fun. Because they are so engaging, they can motivate students to want to understand the meaning of what is being said. According to Sundqvist (2009), the size of the students' vocabulary correlated positively and significantly with the time the students spent on Extramural English activities (Sundqvist, 2009, p. 1).

As part of the Norwegian English curriculum, different core elements are highlighted. These core elements are communication, language learning, and working with texts in English. Concerning language learning, vocabulary is specifically mentioned (Norwegian Directorate for Education and Training, 2019, p. 2). Vocabulary is an integral part of what it means to learn a language. Therefore, it is essential to research how it occurs. It has to be mentioned

that classroom teaching undoubtedly is a source from where students learn new vocabulary. However, it is not the only source "One very active research tradition in the field of second language acquisition (SLA) attempts to establish causal relationships between environmental factors and learning" (Schmidt, 1992, p. 206). The Extramural English these students surround themselves with are significant environmental factors that should be examined as a reason that they might be progressing in the language. This is what this thesis is meant to highlight. Many of the activities are inherently very different, however they are common in some ways. All of the activities allow the students to engage with the language, and some of them contain a substantial amount of authentic English with presumably new vocabulary that the students could pick up.

## 1.2 The aim of this study

This study aims to see if there is a correlation between the time the selected 8th grade Norwegian pupils spend on the different Extramural English activities and the lexical richness of their writing. The main aim of this study is to look at the receptive ways to engage in English activities. The primary focus is on movies, tv-series and audiobooks/radio programs/podcasts. Compared to gaming and chatting, these activities are less participatory for the students that engage in them. They are also interesting because they are not traditionally seen as learning activities, unlike reading. Furthermore, students are engaging in English shows/podcasts and audiobooks, from younger ages than ever before. Therefore, it is essential to measure the effects they can have on students' vocabulary. This study seeks to provide more research on how these activities effect Norwegian students' English vocabulary. To answer the research questions in this thesis, 80 texts from the TRAWL corpus are analyzed, and the results discussed. Three different research questions are the focus of this paper, and these are as follows:

*RQ 1: Does time spent on Extramural English through movies/tv-series, audiobooks/radio programs/podcasts improve the lexical richness of 8th graders?*
*RQ 2: Is there any particular one of these media: movies/tv-series, audiobooks/radio programs/podcasts that show a more significant growth in lexical richness?*
*RQ 3: Does total time spent on Extramural English improve the lexical richness of students?*

# 2.0 Theoretical background

## 2.1 Ways to define a word

When the size of a learner's vocabulary is measured, the focus is usually on how many different words there are (Read, 2000, p. 16). There are different ways of categorizing words when determining vocabulary level. However, it can be challenging to define what a word is (Read, 2000, p. 17). For example, there are brand names that are hard to define because they have no other meaning than the name "Coca-cola, Microsoft, etc". There are also different ways of categorizing words. Some ways to categorize words are tokens, types, lemmas, and word families (Nation, 2001, pp. 7-8). Another way of categorizing words is to count them as flemmas. These different ways of categorizing words are defined in this section.

The number of tokens in a text means that each word is counted, even if it occurs more than once, this is a standard measurement of word counting. Types are unique words which occur more than once but are counted only once. Lemmas are when the headword and some of its inflected forms as well as reduced (n't) forms, are being counted as well, as the same lemma (Nation, 2001, p. 7). The reason why lemma has become an increasingly popular metric for researchers is that these are based on the learning burden of words (Swenson & West, 1934, as cited in Nation, 2001, pp. 7-8). This means that it accounts for the difficulty of learning different words. For example, dance, danced, and dancing, would be counted as different types, but they are not different lemmas. In the Brown Corpus, which is a corpus that contains a computerized lemmatised list (Francis & Kučera, 1982, as cited in Nation, 2001, p. 7), superlative and comparative forms of words are not in the same lemma, meaning they count them separately. Furthermore, "the same form used as different parts of speech (walk as a noun, walk as a verb) are not in the same lemma" (Nation, 2001, p. 7).

There is a way of categorizing words that is similar to lemmas which are called flemmas. A flemma is a modified lemma (Stoeckel et al., 2020, p. 601). This word type contains the word's base form and associated inflection forms (McLean, 2018, p. 823). Flemmas are a grouping of lemmas and the word-family, f + lemma is flemma. The flemmas include all base words for all parts of speech where a word stem is found (Pinchbeck, 2017, p. 32). Therefore, walk as a noun and walk as a verb would be under the same flemma. However, they are not under the same lemma. Simply put, when a headword has multiple parts of speech, inflection

for each part of speech is included in the same flemma (Stoeckel et al., 2020, p. 601). Some research shows that knowing the words in different parts of speech might not be an additional learning burden in all cases. Stoeckel et al., (2018) found that when testing students lexical knowledge in one part of speech, in 56 % of the cases, they knew the same words in different parts of speech (Stoeckel et al., 2020, p. 604). This indicates that flemmas are an even more reliable measurement, as it could better account for the learning burden.

The last relevant categorization of words is word families, which include both inflected forms like lemmas and affixes. When determining what part of a word family is, one may include more prefixes and suffixes to words than when counting lemmas. Some affixes include "- ly, -ness, and un-" (Nation, 2001, p. 8). To put it simply, more variations of the headwords are counted within the same unit than they are when one is looking for lemmas. However, this method of counting words requires the researcher to be explicit in what is or is not part of a word family because this can be subjective to a degree (Nation, 2001, p. 7-8). In this paper word families are not used as a way of categorizing the words the students use. This is because, as mentioned, the word families are somewhat subjective and defining this is beyond the scope of this thesis.

## 2.2. Lexical Richness

There is yet to be a foolproof way to measure a writer's vocabulary. This is because many factors can affect writing, but lexical richness is one way to examine how extensive and varied a writer's vocabulary is. However, this method also has its flaws. Laufer & Nation (1995) point out that factors other than vocabulary could affect lexical richness. They refer to familiarity with a subject, how good of a writer the person is, and the communicative purpose of the text (Laufer & Nation, 1995, p. 308). With this being mentioned, in this thesis, different methods of measuring lexical richness are used, which could help make the results more reliable. The measurements that are utilized are lexical density, lexical sophistication, and lexical diversity. The reasoning for using these three measurements is that they give a fuller view of the vocabulary in comparison to using them alone. While each of the measurements says something about the vocabulary of the students, using them combined gives a better picture of the actual level. In the following sections, these measurements are discussed in more detail.

## 2.2.1 lexical diversity

The first measurement is lexical diversity, which means how many different words the writer uses. This can be measured with the type-token ratio also called TTR. Vocabulary size can be seen as the different word types, compared to the length of a text measured in word tokens (Tweedie & Baayen, 1998, p. 323). The TTR of a text is the number of different word types divided by the number of total words. When using this method, the larger the TTR is, the more varied the vocabulary usage is. "If a speech sample contains 20 words and they are all different, we obtain the ideal TTR: 20/20 = 1.00. On the other hand, a sample in which the same word is repeated 20 times yields a figure of 1/20 = 0.05". (Richards, 1987, p. 201). Ideally, when measuring this, the range of vocabulary should be seen with regard to the length of the texts (Richards, 1987, p. 201). One issue with using only this method of measuring is that the TTR could be high in two very different texts or sentences. One example is two sentences mentioned in Jarvis's (2013) lexical diversity paper.

These are: "1: We run every morning" "2: We run up and down the slope of that hill every morning" (Jarvis, 2013, p. 100). This shows one of the problems with calculating lexical diversity, because both these sentences create a TTR of 1.00. Therefore, they are seen as equally good. Arguably, the second sentence is more complex because it contains more words, but when only looking at the ratio, it is not. Therefore, it is essential to the method that all the texts have a similar length to each other for the results to be reliable. In summary, this method has some disadvantages on its own. However, the inclusion of this measurement alongside the other measurements does help create a fuller picture of the students' lexical richness and is therefore a helpful tool.

Another measurement of lexical diversity that is similar to TTR is the lemma-token ratio, which some researchers argue is a better way of assessing vocabulary (Granger & Wynne, 2000, p. 3). The difference between these two measurements is that the words are divided into lemmas instead of types. When calculating lemma-token ratio, one would categorize the different forms of the same words as just one lemma: jump/jumped/jumping, would be one lemma. In contrast, when using a type-token ratio these words would be three different types (Granger & Wynne, 2000, p. 3). This means that when using lemma-token ratio, one usually gets a lower score than when using TTR. Other researchers believe it to be questionable whether or not lemma-token ratio is more accurate at defining lexical richness than type-token based analysis, and in one study, they found the values to be relatively similar,

especially for speakers with low proficiency (Vermeer, 2000, pp. 78-79). There is also a way to measure the lexical diversity called flemma-token ratio also called FTR, where instead of counting each word as a lemma, one categorizes them as flemmas. Flemma-token ratio is the metric that is used in this paper, alongside type-token ratio, to calculate lexical diversity. Flemmas are similar to lemma, the difference is that flemmas group identical forms of different parts of speech, which lemmas do not.

## 2.2.2 Lexical sophistication

Lexical sophistication also referred to as LS, is another metric that can tell us about the vocabulary level of a text. Lexical sophistication, to put it plainly, is the number of advanced words in a text. There are several ways to measure lexical sophistication, and there is disagreement in the field about which one is the most representative. However, it involves both the depth and breadth of the writers' lexical knowledge (Meara, 1996, 2005a; Read, 1998, as cited in Kyle & Crossley, 2015, p. 759). When measuring lexical sophistication, frequency is a keyword. This is because how advanced a word is, often relates to how frequently the speakers of the language use it. Also, researchers tend to work from the notion that words that are more frequently used in a language are learned earlier than words that are used less frequently (Kyle & Crossley, 2015, p. 759). Therefore, one measures lexical sophistication by finding out how rare the words the writer uses are. Using low-frequency words that are appropriate for what one is writing about, instead of just general vocabulary, increases lexical sophistication (Read, 2000, p. 200). Having more 'advanced' words gives a higher score of lexical sophistication, but what is considered advanced might depend on the researcher's definition (Laufer & Nation, 1995, pp. 309-310). Therefore, it is essential to explicitly express the measurement used for lexical sophistication in a paper. In this thesis, the degree of lexical sophistication is determined by using firstly, The New General Service List corpus released in 2013 and created by researchers Charles Browne, Brent Culligan, and Joseph Phillips (Browne, 2014, p. 1). Secondly, the New Academic Word List, which was also released in 2013 and created by the same researchers (Browne et al., 2013), is used. These corpora contain wordlists, and the less frequent words is seen as more advanced when calculating the score. This method is called determining the lexical frequency profile of the writing. Different tools are available, and these tools allow researchers to utilize these corpora online and make it possible to use this method to give researchers the lexical frequency profile of a text.

### 2.2.3 Lexical density

Lexical density also called LD, is determined by how many content words a text includes. In addition to lexical diversity and lexical sophistication, this last measurement is used as a way to determine lexical richness "Lexical words are the words which primarily convey information, a text is considered 'dense' if it contains many lexical words relative to the total number of words" (Laufer & Nation, 1995, p. 309). This means that lexical density is measured similarly to how TTR is measured, i.e., one looks at the number of lexical words and divide it by the total number of words. Therefore, a score closer to 1.0 would be better when determining a text's lexical density, because it would indicate that the writer uses primarily lexical words. Lexical words are nouns, verbs, adjectives, and adverbs (Johansson, 2008, p. 65). Researchers can use different methods to determine a text's lexical density. For example, one is noun density, where one takes the total number of nouns divided by the number of tokens in the text. This can also be done with verbs, adjectives, and adverbs (Johansson, 2008, p. 65). In this paper, the total number of lexical words is used instead of separating them. There is an issue if one was to determine the level of a text based only on this measurement. This happens in more advanced writing, where one may see an increase in function words because of the use of more subordinate clauses, participle phrases, and ellipsis (Laufer & Nation, 1995, p. 309). This would mean that a text with a lower lexical density could be more advanced and better than a text with a higher lexical density. Nevertheless, with the students level and age in mind, it is likely that the lexical density scores are fairly accurate.

## 2.3 Literature review

This section presents a review of the research field of incidental vocabulary learning for English L2 (second language) learners. Several studies have been conducted on the effect of incidental vocabulary learning, and some of the English activities studied in this thesis (podcasts, movies, tv-series) have been studied independently. As mentioned in Arndt & Woore (2018), learners may learn new words when they see them in a topic that they are engaged in. They then use the context of the word's usage to determine its meaning (Arndt & Woore, 2018, p. 125). In the Arndt & Woore (2018) study, the researchers wanted to see if the incidental vocabulary gain from two different media activities, reading blog posts and watching video blogs, would be comparable. Arndt & Woore tested 84 English foreign language (EFL) learners from different backgrounds to examine this. The participants were

learners with ages ranging from 14 to 25 with 19 different L1s (first languages). The researchers created a post-test, meaning a test given after the exposure to the English input, which contained six pseudo-words so that the students had not encountered the words before, and these were repeated 11-14 times each. Three vloggers created video blogs, and they were free to choose a script. However, they were given instructions by the researchers on how to include the pseudo-words. The pseudo-words were in both the video blogs and the other blogs created from the video scripts (Arndt & Woore, 2018, pp. 128-129). After being exposed to the two types of media, the participants were given the post-test with the pseudo-words. Arndt & Woore (2018) found that reading blog posts and watching video blogs gave a similar amount of incidental vocabulary gain (Arndt & Woore, 2018, p. 135). This study provides evidence that incidental vocabulary gain can occur and that watching videos can have the same effect as reading. This relates to the English activities watching movies and tv-series.

Within this field of research, there are also different ways researchers believe one can maximize vocabulary gain. A central discussion here is the use of L1 subtitles compared to the use of L2 captions with the input. Peters et al's., (2016) study found that vocabulary learning can occur when watching a tv program in class. The researchers created two separate groups. Group number one consisted of 28 higher proficiency Dutch-speaking EFL learners between the age of 17 and 18. 16 of them were put in the L1 subtitles group, and 12 of them in the L2 captions group. Group number two consisted of 18 lower proficiency students with age ranging from 17-20, most of these were also Dutch-speaking EFL learners. 10 of these were in the L1 subtitles group while 8 of them were in the L2 captions group. The researchers gave the students pre-tests one week prior to watching a documentary video that lasted 13 minutes for group one, and a cartoon episode with a 20-minute duration for group two. Both contained a reasonable vocabulary level for the groups' proficiency. They were also given a post-test immediately after watching it (Peters et al., 2016, pp. 141-142), which was where the researchers found that vocabulary learning can occur from these methods of audiovisual engagement. Regarding L1 subtitles compared to L2 captions, they found a difference in the form recall (providing the form of the target item) scores in group two, who watched the cartoons. They found that the L2 caption participants scored slightly better with 21.5% correct answers, and the L1 subtitles participants scored slightly worse with 11.4% correct answers (Peters et al., 2016, p. 143). Peters et al., (2016) mention at the end of their study that they believe more research needs to be conducted on which type of audiovisual input, i.e.,

movies, tv-series, documentaries, or cartoons that would be best suited for vocabulary learning (Peters et al., 2016, p. 146).

In another study on this subject conducted by Aschcroft et al., (2018), the researchers found that it was possible to gain vocabulary just from watching a singular movie with L2 captions. This research was done on 187 native Japanese students from a university in Japan, with ages ranging from 18 to 23 years old (Ashcroft et al., 2018, p. 138). They chose a movie called *Back to the Future* because they wanted to engage students, and based on a survey they did on 43 students from a film class, this was rated as the most enjoyable movie for the students. They also analyzed the script and found 42-word items that were often repeated in the movie (Ashcroft et al., 2018, pp. 138-139). In this experiment, they did pre- and post-vocabulary tests on the students, with the 42-word items. They found that students had a mean gain of 1.77 words, or 4.2% of the total words they were tested in (Ashcroft et al., 2018, p. 144). This offers some evidence that watching movies could be a good way for students to acquire English vocabulary.

In their study conducted in China, Feng & Webb (2020) sought to test the differences between certain modes of input on vocabulary gain, for EFL learners. The participants in the study were 76 students, with ages ranging from 19 to 21. All of the participants were majoring in English translation at a university in China (Feng & Webb, 2020, p. 506). They tested the modes of input by using a television documentary and dividing the students into three groups. The first group read the transcript of the documentary, the second group listened to it, and the last group viewed it. They gave the groups a pre-test before engaging with the documentary and a post-test one week after exposure (Feng & Webb, 2020, p. 499). They found no significant variance between the modes of engagement, and all the modes of engagement yielded some vocabulary learning. This provides further evidence that these activities can promote incidental vocabulary learning (Feng & Webb, 2020, p. 520). All of these studies provide evidence that there is a clear indication that incidental vocabulary learning occurs in students.

### 2.3.1 Extramural English

Tv-series/movies and audiobooks/radio programs/podcasts are the activities that is the main focus of this paper. However, more activities like gaming, reading, chatting, talking, and other specified uses of English are also examined. The students time spent on Extramural

English is gathered through the TRAWL project, where all the students have answered a one-time questionnaire about the time they spent on the certain activities. See the questionnaire in Appendix 1. However, the time spent on activities is not compared with pre or post-tests, but rather singular texts the students have written. The activities are also not part of in class training programs, but rather the time they voluntarily spend on English in their own spare time. These activities are all examples of possible Extramural English activities students engage with. Extramural English is a growing research field, as students these days are spending a lot of their spare time engaging with English content. The studies on this topic often include audiovisual input from either tv shows or movies, reading in English, and sometimes gaming habits of students. Researchers Sundqvist and Sylvén define this exposure as "English outside the walls" (Sundqvist & Sylvén, 2016, p. 6), which means the English that is encountered when learners are not in school. This includes all the activities mentioned. Another criterion is that the English cannot be initiated by the teacher, but rather a voluntary involvement from the student's perspective. It could be that the student engages in English content either because they do it for fun or with the intention of learning. There is a possibility that some students feel pressured to engage in Extramural English activities, and this is also included in the definition as long as it is not instructed by the teacher (Sundqvist & Sylvén, 2016, p. 6). All the activities that are mentioned in this research paper could be viewed as Extramural English as long as it was not instructed by the teacher.

In Sundqvist's (2009) study, the effects of Extramural English were tested. Different Extramural activities were looked at, and their impact on various students' English abilities, among others the effect on vocabulary was examined. 74 students in Swedish schools from the age of 15-16 completed the study (Sundqvist, 2009, pp. 83-85). The students were given two different vocabulary tests, one to test their productive and one to test their receptive vocabulary (Sundqvist, 2009, p. 147). The students were asked to report their time spent on Extramural English activities, and this was analyzed in comparison to the scores on their vocabulary tests. Sundqvist (2009) found a clear connection between the students' time spent on Extramural English and their vocabulary scores (Sundqvist, 2009, p. 154). Furthermore, the importance of the different activities was also looked at with regard to the significance they had in improving the vocabulary of the students. The most impactful activities were 1: playing video games/surfing the internet. 3: watching TV, 4: reading newspapers/magazines, 5: reading books, 6: listening to music, 7: watching films.

Sundqvist (2009) noted that the time spent on "watching films" seemed to be unimportant for vocabulary gain. It is also mentioned that there is a noticeable difference in how much time boys spent on the first two activities and how much time girls spent (Sundqvist, 2009, pp. 156-157). These findings suggest that there is a positive correlation between time spent on Extramural English and vocabulary, but that there is a difference between the activities.

Lastly, with regards to lexical richness, two similar master projects have been conducted using the TRAWL project. The first one is Dasic (2019) and the second one is Auensen (2019). Dasic's (2019) paper sought out to "Investigate the connection between Norwegian lower secondary pupils' gaming habits, their essay grades and their written lexical richness in English" (Dasic, 2019, p. 2). The lexical richness aspect is relevant for this study. In her thesis she ran essays from the students through the Cobb (2023) Vocabprofile program Compleat VP. She found that the differences in the scores from the amount of time the students spent on gaming and their lexical richness to be non-significant (Dasic, 2019, p. 42). Dasic's lexical richness analysis was conducted on 14 texts, and she mentions that an area that could be interesting for further research, was to have a larger sampling size for essays that could provide more meaningful statistical tests (Dasic, 2019, p. 50). This is discussed later in my paper. Auensen's (2019) paper was written about 27 excerpts from Norwegian 9th graders and wanted to "Investigate the Norwegian lower secondary EFL teachers' assessments of their students' written composition" (Auensen, 2019, p. 3). This was done by calculating the lexical richness of the students and checking the lexical richness scores against the grades the students received. The participants were divided into two groups, intermediate with grades 3-4, and high-level with grades 5-6 (Auensen, 2019, p. 3). The thesis found that there was a positive correlation between the grades the students received and the lexical sophistication and lexical diversity scores. However, there was no correlation between their grades and lexical density (Auensen, 2019, p. 50). Auensen (2019) shows that studying ways how students increase lexical richness is relevant for their academic achievements.

### 2.3.2 Audiobooks

When learning a new language, there are many ways that learners can improve their vocabulary. It is well-documented that language input is essential for language learning (Rodgers, 2016, p. 43). Audiobooks are a way to be subjected to a good amount of language input. One of the criteria for incidental vocabulary learning is that the input should contain

context cues and methods of building background knowledge (Rodgers, 2016, p. 43). Also, when the goal is vocabulary acquisition through the input, some language research indicates that students need a high amount of lexical familiarity with the material for it to occur. Different studies estimate it to be between 95% to 98% for vocabulary acquisition (Liu & Nation, 1985; Nation, 2001; Webb & Rodgers, 2009a, 2009b, as cited in Rodgers, 2016, p. 43). Therefore, audiobooks must be at an appropriate level for the students. However, listening to the right audiobooks should be a useful way for students to learn vocabulary incidentally, because they contain a lot of language input.

### 2.3.3 Podcasts and radio programs

Both podcasts and radio programs provide similar learning opportunities because of the input they provide. They allow students to access authentic language input. They are also mostly listened to, and while some podcasts and radio-shows film the casters, it is quite common to just listen to these types of media. Podcasts are a relatively new media form that is becoming increasingly popular, and some students might start listening to them from quite an early age. Podcasts are digital audio and video recordings that are available online, and they can be of assistance in motivating students to learn outside of the classroom (Hasan & Hoon, 2013, p. 131). Podcasts can also become a tool that teachers can use to deliver educational content and help students with learning digitally (Hasan & Hoon, 2013, p. 128). Studies about podcasts effect on language learning are promising, these show that they can help aid in language areas such as pronunciation, vocabulary, grammar, and speaking (Chan et al. 2011; Kaplan-Leiserson, 2005, as cited in Hasan & Hoon, 2013, p. 130). Putman & Kingsley (2009) tested out creating podcasts with specific science vocabulary and had a few different findings. The participants were 58 fifth-grade students. Before providing them with podcasts, the students were given a multiple-choice pre-test with 22 science terms, and they were given a post-test after engaging in classroom instruction with or without the complementary podcasts (Putman & Kingsley, 2009, pp. 103-104). Approximately half of the students had access to the created podcasts, and half of them did not. Putman & Kingsley (2009) found that there was a slightly higher vocabulary gain in the groups that had used podcasts. The researchers also conducted a survey and found that the students had positive attitudes toward the podcasts and agreed that they motivated them to learn vocabulary (Putman & Kingsley, 2009, p. 104). However, as mentioned, these were explicitly designed podcasts rather than something students would be likely to listen to on their own time. General research on the use of podcasts has shown

that podcasts can increase language ability, in particular, the listening skill of users. Also, it could be useful for teachers as another tool for their teaching (Indahsari, 2020, p. 103).

### 2.3.4 Incidental learning and input hypothesis

The notion that incidental vocabulary learning occurs in a second language is widely established, and researchers believe that it can occur when for example, a learner is engaged in extensive reading (Huckin & Coady, 1999, p. 1). Research conducted in this area is often focused on what type and size of vocabulary is necessary to gain new words. It is also focused on how often one needs to be exposed to the words in order to learn them (Huckin & Coady, 1999, p. 1). Researchers who look at how vocabulary is learned, have differing views on what is required to learn new words. Most psychological models that focus on memory, work from the notion that attention is necessary for encoding stimulus into long term memory (Schmidt, 1992, p. 209). In language learning attention can be directed to, for example, form or meaning. Schmidt (1992) argues that even though the intention to learn is not necessary to learn, i.e., when one voluntarily watches an English TV-series where the intent is not to learn, learning can still occur. However, attention to the material is crucial (Schmidt, 1992, p. 209). For incidental vocabulary learning to happen, there is also required a certain vocabulary or word knowledge. In Nation & Kyongho (1995), they found that a 1945-word vocabulary was enough for around 83.4% of the knowledge of the words in a text (Nation & Kyongho, 1995, p. 40). However, according to Laufer (1989), this is not enough for adequate comprehension of a text, which requires a 95% vocabulary knowledge of the words in the text. Laufer's (1989) study showed that in most cases, when the lexical knowledge was under 95%, the comprehension of the participants was impaired (Laufer, 1989, p. 319). In order to understand 95% of the language that a learner would encounter in an average text, it is necessary to know around 5000 words (Laufer, 1989, p. 319). This means that the learner has to know substantially more words to go from 83.4% knowledge to 95%, which is necessary for comprehension. Also, the input received should be based on the level of the student, but if only 95% is required for comprehension, there is room to learn more words in most contexts.

Krashen's input hypothesis must also be mentioned. Briefly explained, the hypothesis is that in order for language learners to progress, they need to receive input from the language. The input has to be at a level that is somewhat higher than the learners' current level. Krashen calls this ($i$+1), $i$ refers to the current competence of the learner (Krashen, 1982, p. 16), and +1 refers to the language slightly above the current competence. A part of this hypothesis is

that this sort of acquisition happens when learners have focused on the meaning, and not on the form of the language they are being subjected to. This part is particularly relevant to Extramural English activities. The reasoning for why this learning occurs is that learners use different mechanisms to understand the language, either context, extra-linguistic information, or knowledge of the world (Krashen, 1982, p. 16).

## 2.4 The Literature review's implications for this study

Based on some of the previously mentioned studies (Arndt & Woore, 2018, Aschcroft et al., 2018, Feng & Webb, 2020, Peters et al., 2016, Putman & Kingsley, 2009, Sundqvist, 2009) there is evidence that engagement with Extramural English could increase vocabulary learning for EFL learners. Specifically, appropriate level English engagement. Dasic's (2019) study did not find any significant correlation between gaming habits and lexical richness scores, but it is mentioned in the paper that this could be because of the low sampling size. The aim of this study is testing this and other Extramural English sources on a larger sample size.

# 3.0 Methodology

The primary research question is, "*Does time spent on Extramural English through movies, tv-series, podcasts, and audiobooks improve the lexical richness of 8th graders?*" In this section of the paper, the chosen methodology is explained. Quantitative methods are employed to answer this question. Some research questions are best explained by how one variable affects another (Creswell & Guetterman, 2021, p. 37). For this paper, the variables that are compared are time spent on the English activities and lexical richness. The TRAWL corpus, now run by the University of Oslo and the University of Agder, is utilized to gather the data for the paper. I am using a large number of the texts that the project has gathered and anonymized. In total, 80 texts from 5-hour tests the students have completed is cut to 250 words and are analyzed. To answer the research questions, I use two different methods of research, surveys and using corpora for text analysis. The participants of this study is 8th grade students from Norwegian schools in the Agder Area. Permission to use the texts and surveys has been granted through the TRAWL project. The texts and the surveys are used as the data collection for the paper. The method of data processing and analyzing is done with the Cobb (2023) Compleat Web VP v2.6 vocabprofile program. Thereafter, it is processed in Microsoft Excel. Excel is used as a way to group the data, where I can also differentiate

between the activities and the levels of the students. The excel sheets contain the information from the surveys, meaning the gender of the pupils as well as different answers from the questionnaire.

## 3.1 TRAWL

In this paper, the TRAWL project is where the student texts and surveys were gathered. The participants of the project have completed a survey about their Extramural English activities. TRAWL stands for Tracking Written Learner Language, and is a corpus filled with authentic school writing that is longitudinal (Hasund et al., 2023, p. 1). The research focus of the group is to develop an understanding of vocabulary, morpho-syntax, syntactic complexity, text organization, and genre and formality level (Dirdal et al., 2021). TRAWL collects texts from a range of different schools, and English texts are collected from years 5-13 (Dirdal et al., 2021).

## 3.2 The methods

One of the methods employed is surveys. The students answered surveys or questionnaires on paper that contains information about their Extramural English habits. The reason this is suitable for this paper is that surveys are a way to simply and effectively collect data (Creswell & Guetterman, 2021, p. 436), making it beneficial for a larger sampling size like this. These answers have been collected from schools, with students as willing participants. The TRAWL surveys also contain information about other factors than just the Extramural English activites, such as country of birth, and if the students have spent time living in an English-speaking country and so on. Because of the scope of this paper, these answers have been excluded, and are not part of the research. The corpora used for text analysis are utilized through the Cobb (2023) vocabprofile program, which contains the two corpora used in this paper NGSL and NAWL. NGSL and NAWL are further explained in section 3.5. Also, the TRAWL texts used are from a corpus of student language, and they are used to determine the different lexical metrics. In summation, the information about time spent on Extramural English is used together with the lexical richness scores of the students' texts, to see if there is a correlation between time spent on Extramural English and their lexical richness scores.

## 3.3 The tests and texts

Laufer & Nation (1995) point out that factors other than vocabulary could affect lexical richness, and one of the factors they mention is the communicative purpose of the text. This relates to the tasks the participants of my study had to answer. So, the differences in tasks could affect the lexical richness of the students. Another factor mentioned in Laufer & Nation (1995) is that familiarity with the topic could affect the lexical richness of the writing (Laufer & Nation, 1995, p. 308). However, the communicative purpose of the texts was similar because the exams were similar. Furthermore, the students should be familiar with the topics as they have the same curriculum. Nevertheless, as an example, one of the exams that some students answered, contained a task where they wrote about two different countries. If the students knew a lot of facts about the countries, they might send in a good paper and get a high lexical richness score. However, some students might be good writers but not very familiar with the topic, therefore, familiarity with the topic might affect these students' scores. The data of this paper come from 4 different end-of-term tests the students had to answer. Most of the texts are from the long answer part of the tests, and they are about writing a story or a text. In one of the classes, most of the texts that were analyzed were from the short answer portion of the test, with the shorter texts including some of their long answer parts. Because the tasks the students answered were different, this could be one of the possible factors that could affect lexical richness.

## 3.4 The website and the vocabprofiler

The website that is utilized is called Lextutor.ca. On this website, there are several tools that are useful with regard to linguistics and specifically for determining the lexical scores of a text. The program used for this paper is a vocabulary profiler which is part of the Compleat Web VP V.2.6 (Cobb, 2023). The website also has information about the different corpora that have been used and gives information about different aspects of a text. This paper specifically focuses on lexical diversity through type-token and flemma-token ratio, lexical density, and lexical sophistication. The scores from these metrics are calculated with the use of the New General Service List and New Academic Word List corpora. The website also allows the user to see the off-list words that have been used in the text (Cobb, 2023), which are the words that do not appear on the NGSL lists or the NAWL lists. The off-list words are examined in the results section, but they are not part of the overall calculation of the lexical sophistication scores of the students.

## 3.5 The wordlists

As previously mentioned in the paper, different corpora are utilized in order to calculate the lexical richness of the student texts. The first word list that is being used is called NGSL. For this paper NGSL 1.01 was used, a new version NGSL 1.2 has since been released. NGSL was made by researchers Brent Culligan, Joseph Phillips, and Charles Browne (Browne, 2014, p. 1). This was done as an update to Michael West's "The General Service List or GSL" compiled in 1953 (GSL; West, 1953). The General Service List was compiled of 2.5 million words of texts and 2000 high-frequency words that were seen as essential for learners of English. As this list is 70 years old this year, there has been a rather substantial shift in the high-frequency words that are common in English. The GSL list contained many terms that were less likely to occur in modern texts. Browne (2014) mentions words such as oar, vessel, spade, cultivator, mill, etc. Therefore, there was a need for a renewal of this list. In the New General Service List or NGSL, Browne and colleagues created a list of around 2800 high-frequency words. By doing this, they vastly expanded the size of the corpus used from 2.5 million words to 273 million words (Browne, 2014, pp. 1-2). Furthermore, they made the NGSL from the most high-frequency words useful for second-language English learners. This was done to cover as high as possible the number of English texts with the least amount of words they could (Browne, 2014, p. 2). In the NGSL they also made the definition of a word more explicit and coherent. Previously, in the theory section, I laid out what counts as a word and different types of words. In the NGSL, the researchers also discuss this. The method they used to distinguish the different words is what is called a modified lexeme approach (Browne, 2014, p. 6). The modified lexeme approach is the same as flemmas and counts words in all forms of speech as the same modified lexeme.

The NAWL list 1.01 is also used to analyze the texts. A new version NAWL 1.2 has since been released, however the NAWL 1.01 corpus has been archived and can be accessed through the link in the literature list. This corpus was created by researcher Browne who was part of the NGSL, and the two other researchers. The intention is for it to be used alongside NGSL to cover more of the language. The NAWL corpus is compiled of 288 million words. 248 million of the words are from texts collected from academic journals, student essays, academic discourse, and non-fiction. 3 million of the words are from MICASE (Michigan corpus of academic spoken), and BASE (British academic spoken English. The last 36 million are from a corpus of the top 100 best-selling academic textbooks (Browne et al.,

2013). Because the words in the NAWL corpus are derived from academic writings, they can be seen as more advanced than the ones in the NGSL list. Using this list can be especially beneficial when it comes to calculating the lexical sophistication of the students, because if they use words from this list, it is advanced for their level. The researchers have compiled the words from the NGSL and NAWL corpora in a way that covers 92% of the language people are likely to encounter when reading in English (Browne et al., 2013). When these lists have been put together, it gives a somewhat reliable holistic view of the text's levels with regard to lexical richness. Even though it does not cover 100% of the language that is being used, 92% can give a very clear indication of how the levels of the students are varied. NGSL and NAWL are essential to this research paper, as it allows for a method of checking the words the students use up against existing lists of words.

## 3.6 Calculating the lexical richness values

In Cobb's (2023) Compleat Web VP v.2.6 program, the text scores with regard to lexical diversity, lexical sophistication, and lexical density are calculated. Previously, in the theory section, I have explained how the different metrics are usually calculated. However, there have been choices that I have made with regards to the program. In this section, I explain the choices and how the program calculates the scores.

### 3.6.1 Lexical density

When looking at lexical density, it is the number of content words that is important (Johansson, 2008, p. 65). Therefore, one has to look at the total number of words in the text and see how many of the total words are content words to find the lexical density. In the program, this is done simply by dividing the number of words by the number of content words. Giving the maximal score of 1.00, which means the text contains 100% content words. In the Cobb (2023) Compleat Web VP v.2.6 program, the lexical density is calculated with regards to the total number of tokens in the text. If there is a text with 250 total tokens and 125 of them are content tokens, the text is given a lexical density score of 0.50. As mentioned, the number of content words is the number of words that are nouns, adjectives, adverbs or verbs (Johansson, 2008, p. 65). Simply put, texts that contain many nouns, adjectives, adverbs and verbs would generally have a high lexical density score.

## 3.6.2 Lexical sophistication

When calculating lexical sophistication, the different corpora mentioned are used. Cobb's (2023) vocabprofile program uses flemmas for the word lists they have. NGSL is divided into three parts. NGSL_1 which contains the 1000 most used flemmas in NGSL, NGSL_2 uses the following 1000 most frequent flemmas, and NGSL_3 uses the last remaining 801 most frequent flemmas. To get the most accurate reading of lexical sophistication, the NGSL_3 has been chosen as a list because it has the most uncommon words compared to NGSL_1 and NGSL_2. Therefore, they are more advanced. The NAWL is also used as these words are also considered more advanced, and this list consists of 963 flemmas. The reasoning for using these lists for this metric is, as mentioned previously, that there is a notion that most words that are more frequently used in a language are learned earlier than words that are used less frequently (Kyle & Crossley, 2015, p. 759). So, the less frequent words are learned later in life, which makes it a sign of good vocabulary if the students can use them from an early age.

The texts have been cut to the same length of 250 words, and I look at both the raw numbers of the lexically sophisticated words and the percentage that they represent of the total number of flemmas in the text. Because the Cobb (2023) program filters out words that could not be flemmas, for example proper nouns, there are some cases where student texts with for example, 4 advanced words, from the NGSL 3 and NAWL list can give different lexical sophistication % scores even if they have the same raw number of advanced words. When I have discussed the different ratios to use for text lexical diversity, I have talked about word types, lemmas and flemmas. However, the most common use to measure the diversity of a text is type-token ratio (Granger & Wynne, 2000, p. 3). This is not because type-token ratio is a better metric but because it was more accessible. Now with the accessibility of programs like the Cobb (2023) vocabprofile it is just as easy to extract the advanced flemmas from the text as the types. Moreover, some researchers argue that "It is probably sensible for pedagogic ends to treat inflected forms of a word as the same type" (McCarthy, 1990, p. 73). Therefore, the lexical sophistication is likely better calculated from the total number of advanced flemmas, which is what I do in this paper.

## 3.6.3 Lexical diversity and FTR VS TTR

When calculating lexical diversity, I have used two different metrics. Flemma-token ratio and type-token ratio, and these metrics are both sensitive to text length. In Koizumi (2012), it is mentioned that when calculating lexical diversity, the effects of text length should be avoided

because it can give misleading results (Koizumi, 2012, p. 60). Also, the TTR for short texts has been shown to be unstable. Specifically, texts with over 200 words were found to be stable, while those below 200 were seen as unstable (Laufer & Nation, 1995, p. 314). Also mentioned in the Laufer & Nation (1995) study is that they cut all the texts from the beginning, meaning the first 300 words were used (Laufer & Nation, 1995, p. 314). Therefore, all of the texts in this project are cut from the beginning to 250 words, using the word count in Cobb's (2023) vocabprofile program.

Type-token ratio is the older of the metrics and has been used for a while to calculate the lexical diversity of texts. As mentioned, the types refer to each different individual word in a text, and the tokens refer to the total number of words (Tweedie & Baayen, 1998, p. 323). This is a rather simplistic way of measuring the vocabulary since there is no nuance to the words being used, and words are counted as different if they are different parts of speech. For example, the verb "LISTS" would be counted as a different word than the noun "LIST" (Browne, 2014, p. 5). So, when counting types knowing both LISTS and LIST means knowing two words. However, the TTR could be useful to include because the score is usually higher than the flemma-token ratio and can show more of a difference. The argument for using the flemma-token ratio is that knowing the inflections of a word or knowing the same word in different parts of speech, is not equal to knowing more words. If this is the case, the flemma-token ratio does give a more accurate depiction of the actual diversity of their language. Even though these are different measurements, both are included as the vocabprofile program allows it. As can be seen in Table 1, there are differences in the scores, and this is to be expected.

## 3.7 Validity and reliability

The validity and reliability of the methods is discussed in this section. When researching lexical richness through text analysis, some choose to shorten the texts so that they are all the same length to make the TTR or FTR more valid. One of the reasons why this is done can be found in the previously mentioned example in section 2.2.1 from Jarvis (2013). Where there are two sentences with the same type-token ratio, and they would even have the same flemma-token ratio. However, these sentences are clearly at a different level, and one is more detailed and advanced, also this shows why it is important to include other metrics than just lexical diversity. The number of words can be a factor when determining lexical richness.

Also, the number of types in a text begins to decline with length, because a longer text would not be meaningful without some repetition of tokens (McCarthy & Jarvis, 2010, p. 382). So, text length does matter with these ratios. A decision was therefore made to make these metrics more accurate, and that was to cut all the texts to 250 words. The full texts largely vary in the number of words used. The longest text had a total of 2154 words, while the shortest text had 279 words. Therefore, the differences in all scores would be affected by this, and primarily lexical diversity. However, cutting the texts does affect the reliability, and some students might receive lower or higher scores than if one were to use the entire text they have written.

In this paper, I use the information from the surveys combined with the information I get from the text analysis to see if there is a correlation between the time students spend on the Extramural English activities and their lexical richness. The surveys the students were given are fairly straight forward. The students answered different questions with regards to age, gender, how many languages they speak, where they are from, and so on. Also, they answered questions about how much time they spent on certain Extramural English activities per week. This was done by checking boxes on paper, with either over 10 hours, 5-10 hours, 1-4 hours, up to 1 hour and no time spent. The surveys were conducted in Norwegian. Since the sample texts and surveys from the participants of this study were from different years, there was slight variation in the surveys the participants answered. However, this is negligible as the formulations were very similar and all the participants have answered the questions relevant for my paper. The most current survey used in TRAWL can be found as Appendix 1. One of the things to mention about the surveys with regard to validity and reliability is that the participants might wrongly estimate the time spent on these activities. Although it is probable that they know roughly how much time they spend on the activities, it could be slightly wrong. Moreover, even if they know how much time they spend, the *self-report bias* has to be considered, Donaldson & Grant-Vallone (2002) mention that people who participate in research tend to want to respond to make themselves look as good as possible. This can lead to them under-reporting behaviors that they see as negative and over-reporting behaviors seen as positive (Donaldson & Grant-Vallone, 2002, p. 247). On the survey, one could argue that there are some activities the students could see as more positive. For example, spending time on audiobooks/podcasts, compared to movies and tv-series. This could lead to some of them overestimating the time spent on one activity while underestimating the time spent on

another. Therefore, when using surveys like this, the *self-report bias* has to be considered as one of the reasons why the research might show slightly misleading results.

Another factor that should be mentioned with regard to drawing conclusions from my findings is that the correlation could go two ways. One thing that could be happening is that the students that spend large amounts of time on the English activities learn vocabulary from spending a substantial amount of time on these activities. Another possibility is that they could already have a wide vocabulary, making it more enjoyable to spend time on the activities. It is also possible that there are additional underlying causes for both engaging with these activities and the students' lexical richness, because correlation does not mean causation. Nevertheless, it is difficult to say how much of the difference comes from these factors, but with the studies mentioned in the literature review, varying degrees of vocabulary learning are possible with these activities. The participants and the generalizability of the study also must be mentioned. This study was conducted on Norwegian 8th graders in the Agder area. Because of the Norwegian context, it can be assumed that the findings would not be the same everywhere. Norwegian students are generally good English speakers and writers compared to most other countries. Therefore, high proficiency in language for the age group should be mentioned. Also, all of these texts were collected from the Agder area. This means that there could be possible differences when compared to other regions of the country.

This study was conducted on 6 classes and 22 students from one school, while the 58 participants were from another. Participation in the TRAWL project is voluntary and based on informed consent. All teachers and students have received detailed information letters about the project, and parents of participants below the age 16 have signed a consent form (Dirdal et al., 2022, p. 118). The ethical approval for this project has been given by the Norwegian Centre for Research Data (Dirdal et al., 2022, p. 118). Also, the texts of the students were anonymized so that there was no information that could lead back to them. In the TRAWL project, all of the names and numbers that could reveal who the student or teacher have been removed (Dirdal et al., 2022, p. 123).

## 3.8 Choices made during the analysis

When going through these texts some choices were made in order to give a more correct picture of the actual level of the students. The first choice was that all the texts were cut to

250 words. To ensure that this did not skew the results, I ran a lexical richness analysis on a few texts, and compared the ratios. They were compared before and after the cut to 250 words. The results are summarized in Table 1:

*Table 1. Lexical richness scores full text and at 250 words.*

| Student codes | Flemma-token ratio | Type-token ratio | Lexical density | Lexical sophistication | Lexical sophistication % |
|---|---|---|---|---|---|
| Full text scores below | | | | | |
| P60337 | 0.27 | 0.32 | 0.42 | 11 | 5.7% |
| P60341 | 0.23 | 0.27 | 0.48 | 14 | 7.7% |
| P60203 | 0.37 | 0.44 | 0.48 | 4 | 3.6% |
| 250-word text scores below | | | | | |
| P60337 | 0.39 | 0.44 | 0.43 | 4 | 4.2% |
| P60341 | 0.36 | 0.39 | 0.47 | 4 | 4.6% |
| P60203 | 0.41 | 0.49 | 0.49 | 2 | 2.1% |

The results show that there is quite a big difference in flemma-token ratio and type-token ratio. There is also a big difference in the lexical sophistication raw numbers. However, the lexical sophistication percentage difference is more negligible. Cutting the texts significantly increases the lexical diversity scores, and slightly decreases the lexical sophistication scores.

There were a few other things that were cut out from the texts as well. The first thing was that I cut all the headers. Many of them said 'Engelsk vår', for example. This is not a spelling mistake and is not part of the answer from the students. Most likely, it is an instruction from their teacher. There is no reason to include this as it would only negatively affect the scores

of the students. Therefore, no texts have headers. Some students wrote "English spring," which would most likely not negatively impact their scores. However, because of the variations, it gives a more accurate picture to cut all of them. Another thing that was cut was the headlines. With the headlines there was also a big difference between the individual students. Some wrote actual headlines, while others only wrote "Task 1" for example. Also, it seemed that quite a few headlines were derived from the tasks the students had and, therefore, might not accurately depict their actual English level. In summary, the first 250 words of the students' answers, not including headers, headlines, and "tasks", were analyzed in this research paper.

## 3.8.1 Grouping of participants

The numbers were all analyzed in Microsoft excel. This required several groupings of the different participants when working with the individual activities. All the groups were made with the idea of each of them representing an answer that could be seen in the survey. None, up to 1 hour, 1-4 hours, 5-10 hours, and more than 10 hours. These were all converted into their averages. None = 0, up to 1 hour = 0.5, 1 - 4 hours = 2.5, 5-10 hours = 7.5, and more than 10 hours = 12.5. When looking at the activities separately, these hour groupings were used as groups, so for this analysis, there were 5 groups the students were put into. The "more than 10 hours" group was calculated with an imagined interval of 10-15 hours for their answers. This could affect the validity of the findings because more than 10 hours could mean a lot more than 15 hours. However, an average had to be made in order to calculate the numbers, and an interval with the same number of hours as the last survey answer is reasonable. When calculating the effects of multiple activities combined, a system was created to group the students. The way this was done is that all the hours from the activities combined were added together. When calculating the effects of for example, tv-series with and without subtitles and audiobook/radio programs/podcasts, one of the students can be used as an example; student *P60500* spent 2.5 hours on TV with subtitles, 2.5 hours on TV without subtitles, and 0 hours listening to audiobooks/radio programs/podcasts. Therefore, this student was given a score of 5, meaning 5 hours in total on these activities. After calculating all the students' total number of hours spent on these five activities, I organized the number of hours spent from low to high. I divided the 80 total students into three (low, moderate, and high). Because 80 divided by three is 27 if rounded up, the 27 first students with the lowest number of hours spent were put in the low category. The highest number here was 5 hours spent. Afterward, I looked at the highest number for the next 27 students, which was 7.5

hours. Therefore, the moderate level contained students from 5.25 hours to 7.5 hours. Also, a high level was created. The high level consisted of the students who spent more than 7.5 hours on the activities combined and were seen as the high level of engagement. This same method of grouping was also done when calculating the total level of Extramural English hours, and the students were also put into the categories (low, moderate, high). However, when creating the categories for all time spent on Extramural English, the cutoff for categories was 26 students instead of 27. This was done because two of the participants had to be removed when calculating this. The reason for removing the two participants is explained in section *4.3*.

Some students did not answer all the questions. The students who had "N/A" (not available) in all their categories were removed from the analysis. Some of the students had "N/A" on one or more categories, while they had answered the others. The decision was made to include these answers still and put "N/A" together with "None" as a score of 0, meaning that they do not spend time on the activities. One can reasonably assume that the students did not answer the question because they did not spend time on the activities. However, this effects the findings' validity because one cannot be sure of the students' intention. Furthermore, a student answered between two categories, for example, "5-10 hours (7.5)/1-4 hours (2.5)". Here the averages were calculated, and they were given a score of 5 hours. The same student also answered, "up to 1 hour (0.5)/none (0)" They were given a score of 0.25 hours spent on the given activity. They are included in a separate section in the tables when looking at the activities on their own.

As mentioned, the findings are based on 4 different end of term tests the students completed. Two of the tests were 1-day tests, while in the two others, the students had 2 days to answer. For the texts that had 2-day answers, there were a V1 and V2 (version 1 and 2) that the students submitted. The V1 was selected for the texts, because using the V2 would give these students an unfair advantage compared to those who only submitted 1 version, because they would have had more time to think about the tasks. The V1/V2 groups were also given feedback on their first submission, which would also give even more of an advantage. Because of these factors the V1 was chosen, since using the V2 would damage the validity of the findings.

# 4.0 Results from the analysis

In this section of the paper, the lexical richness analysis results from the Cobb's (2023) program in relation to the Extramural English activities the students spend time on are presented. There are a few different things that are examined in this section. Firstly, I analyze some of the activities individually. This includes time spent on tv-series/movies, here I also see if subtitles make a difference in the scores. I also look at podcasts/audiobooks/radio programs individually to see if there is a correlation. Because of the content of the surveys the students have completed, there is also more information that I go through, like gender differences in time spent on the Extramural English activities. Later, I include all the activities that the students can spend time on (reading online, reading books/newspapers/magazines, playing videogames, chatting in English, talking in English, tv-series/movies, audiobooks/radio programs/podcasts), and other activities that the students might use. This is done in order to see if the combined time spent on these activities makes a difference in lexical diversity, lexical density, and lexical sophistication. Furthermore, I review the off-list words the students have used. Lastly, I examine if the time spent on Extramural English effects the total number of words the students use. I review the group sizes from the different activities and if any activities, in particular, stand out.

## 4.1 The activities on their own

In this section, the research question RQ 2 *"Is there any particular one of these media: movies/tv-series, audiobooks/radio programs/podcasts that show a more significant growth in lexical richness?"* is the main focus.

### 4.1.1 Audiobooks/radio programs/podcasts

Here, the effect on lexical richness from audiobooks/radio programs/podcasts is reviewed. The participants have been divided into 5 groups as seen in 3.8.1, and the average (mean) of the different lexical richness metrics have been calculated.

*Table 2. Lexical richness scores for audiobooks/radio programs/podcasts usage.*

| Audio/radio/ podcast | Flemma - token ratio (mean) | Type-token ratio (mean) | Lexical density (mean) | Lexical sophistication (mean) | Lexical sophistication in % (mean) |
|---|---|---|---|---|---|
| Group 1 (0 hours) (64 participants | 0.409 | 0.476 | 0.462 | 5.05 | 4.86% |
| Group 2 (0.5 hours) 9 participants | 0.420 | 0.490 | 0.470 | 5.44 | 5.09% |
| Group 3 (2.5 hours) 4 participants | 0.415 | 0.474 | 0.468 | 5.15 | 4.89% |
| Group 4 (7.5 hours) 1 participant | 0.410 | 0.500 | 0.460 | 6 | 6.00% |
| Group 5 (12.5 hours) 2 participants | 0.415 | 0.490 | 0.455 | 5.27 | 5.00% |

These are the results from the text analysis with regard to time spent on audiobooks/radio programs/podcasts. A few things must be mentioned here. Firstly, there is a large difference in group sizes. Group 1 contains almost the entirety the participants, and group 4 only contains 1 participant. Therefore, in terms of reliability, these group sizes are far from ideal. All of the ratios in all of the groups are fairly similar. Group 4 is the highest scoring group, but as mentioned, this is only 1 participant, and therefore no conclusions can be made.

## 4.1.2 Tv-series/ Movies with Norwegian subtitles

The table presented below shows the scores from the different measurements in relation to the time spent on Tv/movies with Norwegian subtitles.

*Table 3. Lexical richness scores for tv/movies with Norwegian subtitles.*

| Tv/movies Norwegian subtitles | Flemma-token ratio (mean) | Type-token ratio (mean) | Lexical density (mean) | Lexical sophistication (mean) | Lexical sophistication % (mean) |
|---|---|---|---|---|---|
| Group 1 (0 hours) 4 participants | 0.410 | 0.477 | 0.461 | 5.18 | 4.96% |
| Group 2 (0.5 hours) 10 participants | 0.411 | 0.477 | 0.462 | 5.25 | 5.01% |
| Group 3 (2.5 hours) 41 participants | 0.410 | 0.476 | 0.462 | 5.06 | 4.85% |
| Group 4 (7.5 hours) 19 participants | 0.410 | 0.476 | 0.462 | 5.15 | 4.96% |
| Group 5 (12.5 hours) 5 participants | 0.418 | 0.483 | 0.468 | 5.62 | 5.28% |
| Student with 5 hours spent | 0.430 | 0.520 | 0.520 | 4 | 4% |

Looking at this table, there are a few interesting things. The groups here are much more evenly divided than with the podcast group. The results show that most of the participants spend between 2.5 to 7.5 hours each week on this activity. However, there are very few differences in the scores. The only group worth mentioning is group number 5, where there is

a slight increase in all of the scores, most notably the difference in lexical sophistication. Nevertheless, this group is tiny compared to the total number of participants, and it is therefore not possible to draw any conclusions from this.

### 4.1.3 Tv-series/ movies without Norwegian subtitles

In Table 4 below it can be seen how the time spent on Tv/movies without Norwegian subtitles effect the lexical richness of the 8th grade students.

*Table 4. Lexical richness scores for tv/movies without Norwegian subtitles.*

| Tv/movies without Norwegian subtitles | Flemma-token ratio (mean) | Type-token ratio (mean) | Lexical density (mean) | Lexical sophistication (mean) | Lexical sophistication % (mean) |
|---|---|---|---|---|---|
| Group 1 (0 hours) 23 participants | 0.410 | 0.477 | 0.462 | 5.11 | 4.90% |
| Group 2 (0.5 hours) 15 participants | 0.410 | 0.477 | 0.462 | 5.10 | 4.89% |
| Group 3 (2.5 hours) 26 participants | 0.412 | 0.479 | 0.464 | 5.26 | 5.02% |
| Group 4 (7.5 hours) 10 participants | 0.411 | 0.478 | 0.464 | 5.22 | 5.01% |
| Group 5 (12.5 hours) 5 participants | 0.414 | 0.479 | 0.461 | 5.33 | 5.05% |
| Student with 0.25 hours | 0.430 | 0.520 | 0.520 | 4 | 4% |

The students are fairly evenly distributed between the groups. However, the students spend more time on tv-series/movies with subtitles than they do on tv-series/movies without subtitles. Of the major groups, group 5 again has the highest score in lexical sophistication. Lexical sophistication is where the most noticeable difference has been in all the activities. This table is very similar to the one where the students watch with subtitles, and there seem to be no major differences in the scores.

Regarding the research question, "Is there any particular one of these media that show a more significant growth in lexical richness?" it seems that higher engagement in these activities does not mean that the students acquire a higher degree of lexical richness. All of the results are very similar, and if there are differences, they are more likely to be attributed to group size than the actual engagement of the students. The findings are further discussed in the discussion section of this thesis. Lastly, a scatterplot with the data from the tv/movies with and without subtitles was created with ggplot2 package in R (Field et al., 2012, pp. 136-140, Wickham, 2016). The flemma-token ratio, lexical density and lexical sophistication scores are included to visualize the results. This was to see if there was a normal distribution, which there was not. The scatterplots are shown in the figures 1, 2 and 3.

**Figure 1**

*Flemma-token ratio in relation to hours spent on movies/tv-series*

**Figure 2**

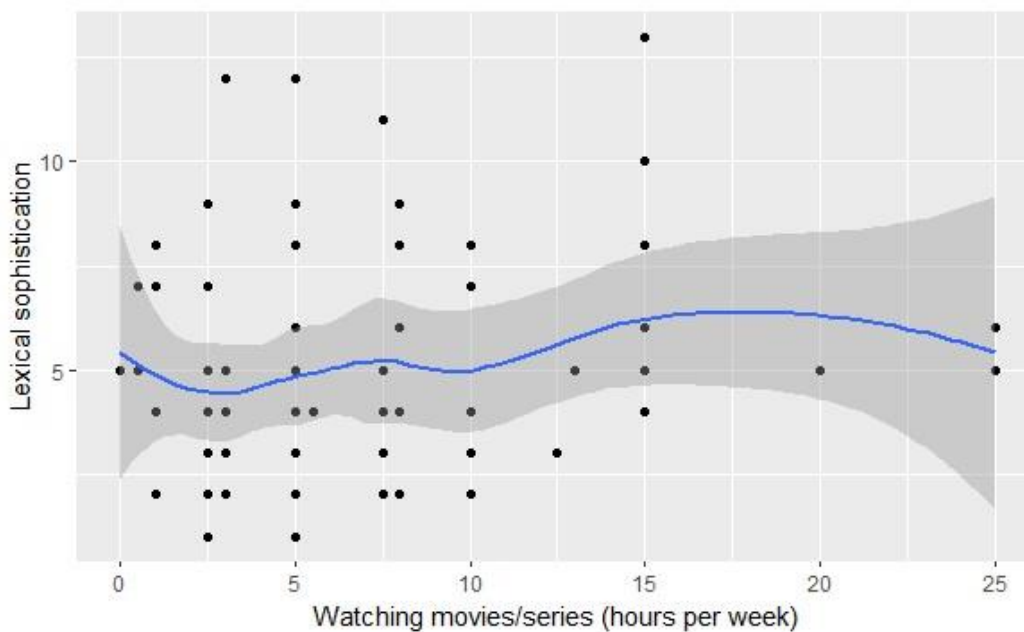*Lexical density score in relation to hours spent on movies/tv-series*
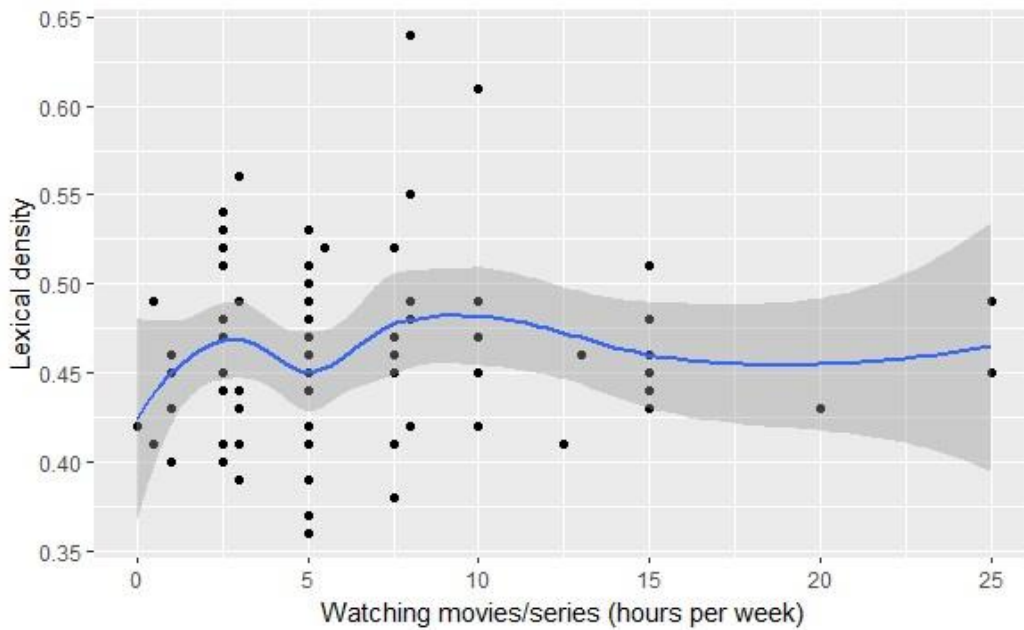


**Figure 3**

*Lexical sophistication score in relation to hours spent on movies/tv-series*



As one can see from these scatterplots the results are not correlated to the hours spent. There are some students who spend no time on the activities that get a high score. While others who spend a lot more time have lower scores at some points.

## 4.2 Time spent on tv-series/movies combined with time spent on audiobooks/radio programs/podcasts.

In this section, the focus is on the first research question I have posed "*Does time spent on Extramural English through movies/tv-series, audiobooks/radio programs/podcasts improve the lexical richness of 8th graders*?" I have created a system where I have divided the pupils into three different groups, using the mean time spent on all these activities combined. With this system, the low engagement users are those who engage 5 hours or below. The moderate are those who engage between 5.25 and 7.5 hours. The high-engagement pupils are those who engaged for 7.5 hours and over. The high-engagement group had a wide range of engagement, and some students spent up to 37.5 hours on these activities combined. In Table 5 the students are grouped into low, moderate, or high based on how much time they spend.

*Table 5. Lexical richness scores for tv/movies, audiobooks/radio programs/podcasts combined.*

| Tv/movies, audiobooks, radio programs and podcasts combined | Flemma-token ratio (mean) | Type-token ratio (mean) | Lexical density (mean) | Lexical sophistication (mean) | Lexical sophistication % (mean) |
|---|---|---|---|---|---|
| Low (5 hours and under) 45 participants | 0.410 | 0.476 | 0.462 | 5.07 | 4.86% |
| Moderate (5.25-7.5 hours) 11 participants | 0.411 | 0.476 | 0.462 | 5.17 | 4.95% |
| High (over 7.5 hours) 24 participants | 0.411 | 0.478 | 0.464 | 5.22 | 5.01% |

There is a slight difference in lexical scores between the groups. However, the difference is minimal and most likely a result of chance. The difference is the highest in lexical sophistication. It could be the case that spending that amount of time on the activities slightly increases the number of advanced words the students know. However, the gap from the moderate to the high group is only 0.05, meaning that the difference is negligible. It does not appear that the time spent on these activities increases the overall lexical richness of the students in any meaningful way. Regarding the groupings, it is clear that most of them are low engagers in these activities, but there are also many in the high engagement group. Also, it is worth noting that the differences across the groups were sizable here on time spent. Some students spend 0 hours while others spend 37.5 hours.

## 4.3 Total time on Extramural English effect on lexical richness and number of words

In this section, the "*Does total time spent on Extramural English improve the lexical richness of students?"* research question is analyzed. Here all the activities mentioned in the questionnaire have been combined, and the total number of hours spent on Extramural English is calculated, even including the questions about other specified types of English use. The other types of activities the students engaged with included "Instagram, Snapchat, and chatting with family members in the US". Two of the students' scores were removed here. This was because they answered that they spent "1-4 weeks" a week on reading English on the internet. One could argue that they most likely meant 1-4 hours. However, the previous question related to time spent living in an English-speaking country. Furthermore, as there is no way to say for sure what they meant, the texts were removed from consideration. Therefore, there were 78 participants in total here. Also, I have calculated the time they spend on average on each of the Extramural English activities. These are as follows:

Reading English on the internet: 2.89 hours

Reading English from books/newspapers/magazines: 0.65 hours

Playing computer games in English: 3.18 hours

Chatting, writing emails/sms in English: 0.96 hours

Talking to someone in English: 0.98 hours

Watching series/movies with Norwegian subtitles: 3.94 hours

Watching series/movies without Norwegian subtitles: 2.56 hours

Listening to audiobooks/radio programs/podcasts in English: 0.60 hours

Other specified uses of English: 0.45 hours

Total mean time spent: 16.21 hours

*Table 6. Lexical richness scores from the total time spent on Extramural English.*

| Total hours on Extramural English combined | Flemma -token ratio (mean) | Type- token ratio (mean) | Lexical density (mean) | Lexical sophisticati on (mean) | Lexical sophisticati on % (mean) | Total number of words (mean) |
|---|---|---|---|---|---|---|
| Low - 8 and under hours (27 participants) 19 female 8 male | 0.397 | 0.463 | 0.454 | 4.62 | 4.58% | 694 |
| Moderate -8.5- 16 hours (25 participants 16 female 9 male | 0.408 | 0.469 | 0.466 | 4.72 | 4.58% | 671 |
| High - over 16 hours (26 participants) 8 females 18 males | 0.423 | 0.495 | 0.465 | 5.84 | 5.41% | 731 |

When the total number of hours was calculated, the three-way division became close to equal for the low, moderate and high group, with 27, 25 and 26 participants, respectively. This allows for a more accurate result from each group. A vast majority spend between 0 and 16 hours a week on Extramural English (52 out of 78 of the participants). However, there are some extreme outliers who report that they spent 73 and 82.5 hours a week on the English activities combined. The first thing to note is that here there are a few differences in most of the categories. The high group has better scores in all categories compared to the other

groups. Again, as with the other tables, the difference is most substantial in the lexical sophistication category. With regards to total words, there is also a difference in the groups, but the low category has more total words than the moderate category, while the high category has the most total number of words on average. It should also be mentioned that there was a substantial difference between the number of words in the shortest text, which was 279 words, and the longest text which contained 2154 words.

Scatterplot graphs were created here as well using ggplot2 in R (Wickham, 2016). This graph included flemma-token ratio, lexical density and lexical sophistication. These graphs also show a more random distribution and provide further evidence that the activities did not improve the lexical richness scores of students. The correlation between the time spent on the different extramural activities and the complexity of learner texts was tested with the *cor*() function in R (R Core Team, 2022) using Spearman's correlation coefficient because several of the variables were not normally distributed (Field et al., 2012, pp. 223-225). The Spearman correlation analysis did not show any sizeable correlation between the time spent on the activities and the lexical complexity of the students texts, this was according to the metrics of the test, where 0.1 is a small effect, 0,3 is a medium effect, and 0.5 is a large effect (Field et al., 2012, p. 58). For both the scatterplot and the Spearman test, all N/A texts were removed, and also the student that answered between categories. so, the sample consisted of the *77* remaining participants. The scatterplots and the tables from the spearman tests can be seen below.

**Figure 4**

*Flemma-token ratio in relation to hours spent on Extramural English*
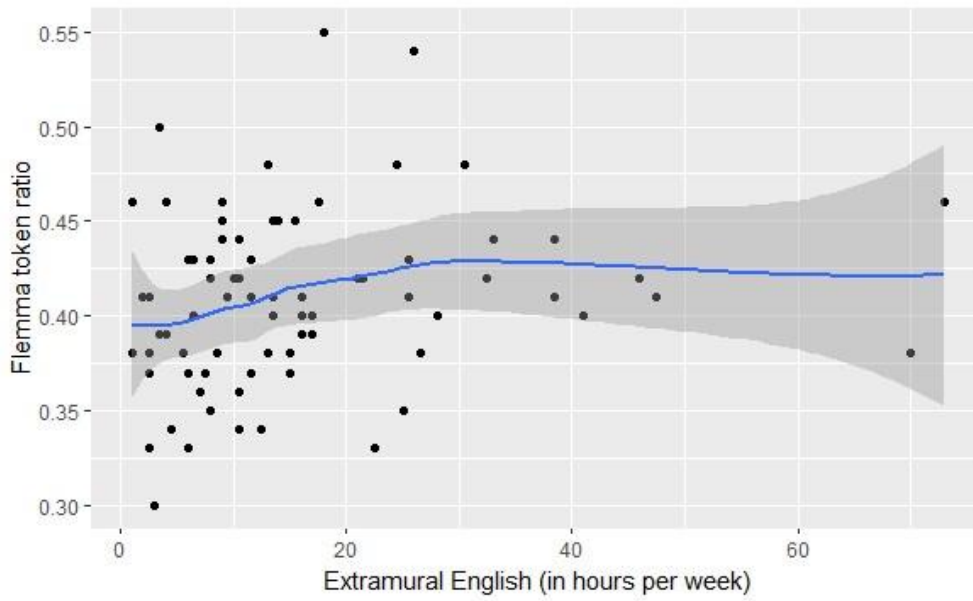


**Figure 5**

*Lexical density in relation to hours spent on Extramural English*
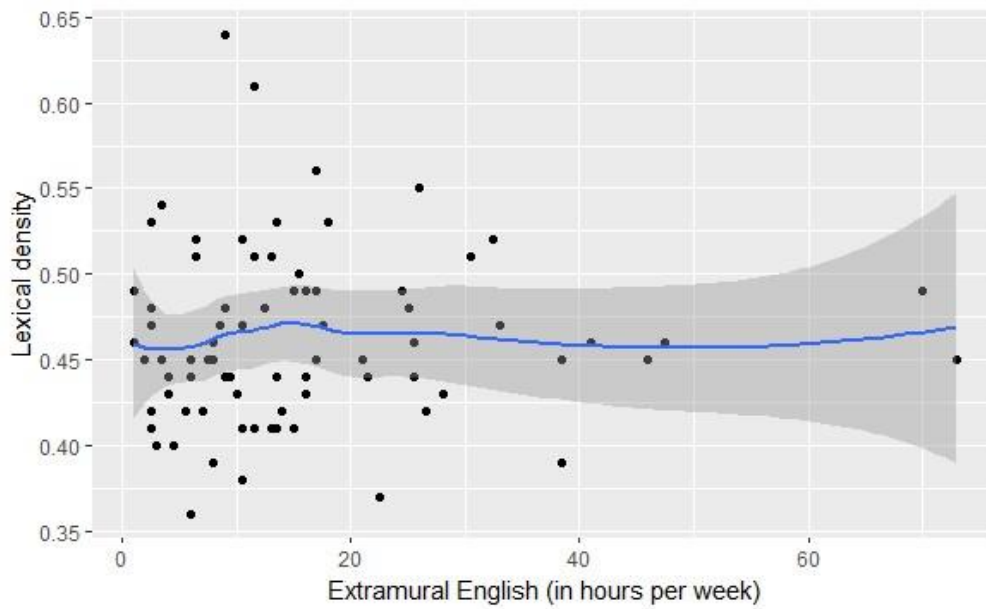
**Figure 6**

*Lexical sophistication in relation to hours spent on Extramural English*
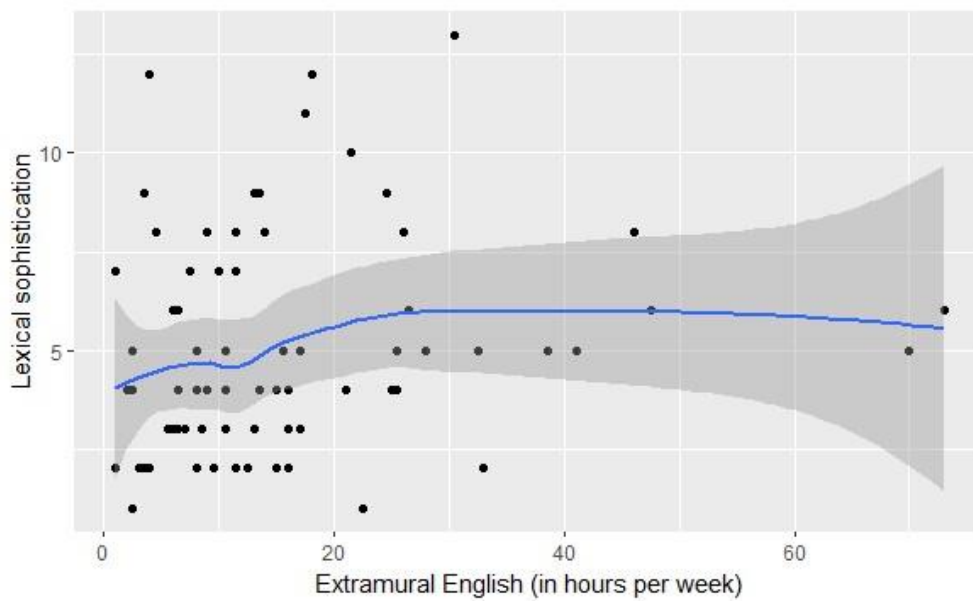


*Table 7. Spearman's correlation coefficients.*

|  | Total time | Flemma-token ratio | Lexical density | Lexical sophistication | Lexical soph percent |
|---|---|---|---|---|---|
| Total time | 1.00 | 0.228 | 0.096 | 0.242 | 0.225 |
| Flemma-token ratio | 0.228 | 1.00 | 0.427 | 0.566 | 0.409 |
| Lexical density | 0.096 | 0.427 | 1.00 | 0.199 | 0.135 |
| Lexical soph | 0.242 | 0.566 | 0.199 | 1.00 | 0.976 |
| Lexical soph percent | 0.225 | 0.409 | 0.135 | 0.976 | 1.00 |

When examining Table 7 it is clear that there is only a small correlation between total time and flemma-token ratio and lexical sophistication with ratios of under 0.3. With the lexical density the score is under 0.1 which means there is no correlation.

*Table 8. Activities seperately with the metrics.*

| | Reading online | Reading books | Games | Chatting | Talking | Watching sub | Watching no sub | Listening | Flemma-token ratio | Lexical density | Lexical sophistication | Total words |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reading online | 1.00 | 0.258 | 0.550 | 0.561 | 0.376 | 0.320 | 0.277 | 0.125 | 0.257 | -0.026 | 0.345 | -0.057 |
| Reading books | 0.258 | 1.00 | 0.160 | 0.370 | 0.371 | 0.102 | 0.276 | 0.154 | 0.184 | 0.078 | 0.013 | 0.099 |
| Games | 0.550 | 0.160 | 1.00 | 0.383 | 0.564 | 0.116 | 0.249 | 0.189 | 0.093 | 0.092 | 0.199 | -0.204 |
| Chatting | 0.561 | 0.370 | 0.383 | 1.00 | 0.409 | 0.213 | 0.391 | 0.116 | 0.140 | 0.013 | 0.202 | 0.096 |
| Talking | 0.376 | 0.371 | 0.564 | 0.410 | 1.00 | 0.203 | 0.234 | 0.313 | 0.054 | 0.087 | 0.043 | -0.146 |
| Watching sub | 0.320 | 0.102 | 0.116 | 0.213 | 0.203 | 1.00 | 0.079 | 0.193 | 0.055 | 0.133 | 0.064 | -0.120 |
| Watching no sub | 0.277 | 0.276 | 0.249 | 0.391 | 0.234 | 0.079 | 1.00 | 0.295 | 0.104 | -0.031 | 0.187 | -0.118 |
| Listening | 0.125 | 0.154 | 0.189 | 0.116 | 0.313 | 0.193 | 0.295 | 1.00 | 0.103 | 0.054 | 0.135 | -0.148 |
| Flemma-token ratio | 0.257 | 0.184 | 0.093 | 0.140 | 0.054 | 0.055 | 0.104 | 0.103 | 1.00 | 0.427 | 0.566 | 0.140 |
| Lexical density | -0.026 | 0.078 | 0.092 | 0.014 | 0.087 | 0.133 | -0.031 | 0.054 | 0.427 | 1.00 | 0.199 | 0.111 |
| Lexical sophistication | 0.345 | 0.013 | 0.199 | 0.202 | 0.043 | 0.064 | 0.187 | 0.135 | 0.566 | 0.199 | 1.00 | 0.125 |
| Total words | -0.057 | 0.099 | -0.204 | 0.096 | -0.146 | -0.120 | -0.118 | -0.148 | 0.140 | 0.111 | 0.125 | 1.00 |

In Table 8 it is notable that all the activities are still under 0.3 in correlation in relation to the lexical richness metrics, meaning there is only a small correlation between some of them, and some metrics have a negative correlation. As can be seen, some of the variables have a small effect on each other, but there is too much variance and too little effect to draw any meaningful conclusions from this. The table also shows that there is some correlation between the activities in relation to each other, i.e., reading online and gaming. However, this is not discussed further as it is beyond the scope of this paper.

### 4.3.1 Gender division between categories

As mentioned previously, the low, moderate, and high categories have a very similar number of participants in each bracket, when looking at total time spent on Extramural English. However, there is a substantial difference in the number of participants from each gender in the brackets. The low and moderate groups have significantly more participants from the female gender. Also, there are more females participating in total. There are 43 female and 35 male participants. Nevertheless, the high group has by far the most males, and it contains over 50% of the total male participants. This finding indicates that there is a big difference in how much each of the genders spends on the activities. In fact, 9 of the 10 participants that spent the most time (between 31 and 82,5 hours) were male.

*Table 9. Gender division in time spent on activities.*

| Activities | Male (35 participants) mean hours | Female (43 participants) mean hours |
|---|---|---|
| Reading English on the internet | 3.01 | 2.22 |
| Reading English from books/newspapers/magazines | 0.68 | 0.31 |
| Playing computer games with English | 3.34 | 2.32 |
| Chatting, writing emails/sms in English | 1.02 | 0.61 |
| Talking to someone in English | 1.04 | 0.51 |
| Watching series/movies with Norwegian subtitles | 4.10 | 3.58 |
| Watching series/movies without Norwegian subtitles | 2.69 | 1.98 |
| Listening to audiobooks/radio programs/podcasts in English | 0.63 | 0.19 |
| Other specified uses of English | 0.47 | 0.22 |
| Total hours mean time of all activities added together | 17.03 | 11.99 |

As shown on the table, there is quite a sizeable gap in how many hours the genders spend on Extramural English. The male participants spend more time on all the activities when looked at individually. They particularly spend more time on gaming than the female participants, where they spend over 1 hour more a week. However, all 8 participants that spent 12.5 hours on average on gaming were male, which raises the average.

## 4.4 Off-List words findings

Firstly, it must be mentioned that the off-list words were counted as *types*. The off-list words are the words that do not appear on the NAWL or NGSL lists. Even though the off-list words that the program show were not used in calculating the lexical sophistication, they were still noted and included in the excel sheet. The highest number of off-list words in a text was 40, but the off-list words in this text were mostly spelling mistakes. The lowest number of off-list words in a text was 0, and only applied to one text. The average number of off-list words in a text was 7.81 words. The most common advanced word used was "bullying" and variations of it, which is mentioned in the "you can sit with me" task some students had to answer for the exam. Nevertheless, some of the words were advanced for the students age, such as "hemisphere, indescribable, awkward and dungeon" and did not seem to be directly attributed to the task texts. Because of the scope of this paper the lexical richness scores were not calculated using the off-list words that were correctly spelled.

# 5.0 Discussion

In this section of the paper, the findings from the study are discussed. This is done in the same order as they are posed in the first chapter of the paper, and I offer my thoughts on the data.

## 5.1 RQ 1: *Does time spent on Extramural English through movies/tv-series, audiobooks/radio programs/podcasts improve the lexical richness of 8th graders?*

This is the first research question that is presented. The total time the students spent on these five activities were added together and compared to the lexical richness metrics that were used in this paper. The results showed that there was no meaningful correlation between the time the students spent on these activities and their lexical richness scores. There could be several reasons for why this occurred. The first reason that should be looked at is the types of

activities. Sundqvist (2009) mentions some Extramural English activities as more effective in showing vocabulary gain than others. Watching tv is on place 3/7 of the most useful activities for vocabulary gain, and watching movies ranks last at 7/7 (Sundqvist, 2009, p. 156). Specifically designed podcasts were shown by Putman & Kingsley (2009) to yield a slight vocabulary gain for students. Audiobooks and radio programs are also ways for the students to be subjected to a lot of authentic English. All of these activities (movies/tv-series/audiobooks/radio programs/podcasts) are ways to engage with the language for the students. There may be differences between the levels of input the students engage with, but vocabulary gain should be possible from all of them. While it seems from the results that there is no relationship between the time spent on these activities and the lexical richness of the students, there could be other reasons why this was the case. As mentioned, there is a difference between the effectiveness of the activities. There might also be a difference in the type of input they receive. As stated in Krashen's (1982) input hypothesis, the input level has to be the level they are at and some slightly more advanced words ($i+1$). Therefore, it is possible that the activities are not at an appropriate level for each student. This would be an interesting area for further research, where one could have more information about specifically what the student watches to determine the input level they receive.

When analyzing the texts in the program, there is also a possibility that the tasks could have played a role. As mentioned, there were 4 different tests handed out with different tasks on them. Some of the texts analyzed were mostly short answers, which could also affect the scores of the different metrics. A task about writing a story could have more adjectives which would affect lexical density than a more logical information-based task. Furthermore, there was a difference in group sizing when combining these activities. Creating many participants in both the low (45) and high (24) groups. However, this might not have made a difference since the scores are so similar. Lastly, as mentioned, the texts have been cut to 250 words. This was done because, as Koizumi (2012) mentions, different text lengths can give misleading results with regard to lexical diversity. As shown in Table 1, there was quite a big difference in both FTR and TTR when cutting the texts. However, one issue with this is that after the cut, some of the students only answered one short task, while others answered up to three tasks within this word count. Therefore, some have written in a few different genres, which could affect the lexical richness scores they receive. It might have been a good idea to create one task for all participants with a word count of 250 words. This way, some of the factors that affect the reliability of the research could be limited.

## 5.2 RQ 2: *Is there any particular one of these media: movies/tv-series, audiobooks/radio programs/podcasts that show a more significant growth in lexical richness?*

In this section, I discuss the findings from each individual activity. Here I compare the scores, but I have decided not to use the score of the single pupil that answered between categories, and I focus on the activities separately. As mentioned, tv-series/movies are grouped, as well as audiobooks/radio programs/podcasts. The survey differentiates between the movies/tv-series with or without subtitles. Firstly, the different scores in lexical richness for movies/tv-series with or without subtitles are discussed. Peters et al's., (2016) study is relevant here, where the researchers found that there was a vocabulary gain from engaging with 13-20 minute audiovisual input from tv, and a slightly higher vocabulary gain on form recall from using L2 captions, compared to L1 subtitles (Peters et al., 2016, p. 143). In the TRAWL questionnaire, L2 captions are not mentioned, only L1 subtitles or no subtitles. In my findings, there was not a mentionable difference between time spent watching movies/tv-series with L1 subtitles compared to not using subtitles. For the time spent engaging in tv-series/movies with L1 subtitles, the scores ranged from: FTR: 0.410-0.418, TTR: 0.476-0.483, LD: 0.461-0.468, LS: 5.06-5.62, LS%: 4.85%-5.28%. The scores from the time spent watching tv-series/movies without subtitles ranged from: FTR: 0.410-0.414, TTR 0.477-0.479, LD 0.461-0.464, LS 5.11-5.33, LS %: 4.89%-5.05%. There is not a meaningful difference concerning any of the metrics of lexical richness. Peters et al., (2016) mention that they had checked the learning materials they used (a 13-minute documentary and a 20-minute cartoon episode) and that they were seen as a suitable level for the groups. The material the students engage in could be a factor in how much vocabulary gain they get. Arndt & Woore (2018) also concluded that it was possible to gain vocabulary from watching L2 audiovisual input. However, the material used was specifically created for the students. Putman & Kingsley (2009) also created podcasts for the students, and they had the intent to teach vocabulary with their podcasts. Ashcroft et als'., (2018) input was a single movie called *Back to the Future. Back to the Future* is a normal Hollywood movie and likely to be somewhat comparable in level to some of the tv-series or movies that the students in this study watch. They found, as mentioned, a 4.2% mean gain in vocabulary from watching the movie. In the Feng & Webb (2020) study they found that there was vocabulary learning from watching a documentary they showed. The vocabulary gain was comparable to the other modes of input, which were reading and listening.

There are differences that could explain why these studies found a gain in vocabulary that was not present in my own findings. The first difference is the method of evaluating the participants vocabulary. Several of the mentioned studies conducted vocabulary tests, I measured the lexical richness by analyzing written answers to exam questions. Another difference is the age of the participants. All of the students from the studies were from the age of 14-25, with the exception of the Putman & Kingsley (2009) study that was done on 5th graders. Therefore, the findings they had could have been different in another age group, like the 13/14-year-olds that I base my study on. However, it is likely that in their own time, students engage with input meant for their age group, and they should, based on the studies reviewed, pick up vocabulary from this. One more difference is that the material was either chosen or created in the studies mentioned in this section. When students engage on their own, it is hard to know what they are watching, and the input could be extremely varied from student to student. However, it is difficult to say whether input and age effects the findings to such an extent as was shown in my analysis. It is also likely that the main difference between these studies and my research is the method of data collection and analysis. Therefore, no conclusions can be made with regards to these studies compared to mine, the differences can only be pointed out.

In the *3.7 validity and reliability* section Donaldson & Grant-Vallone (2002) *self-reporting bias* was discussed. The *self-reporting bias* could mean that in the surveys students would overestimate the time they spent on the activities seen as positive, while underestimating the time on the activities they see as more negative. Here I used an example that students might view it as more agreeable to spend time on audiobooks/radio programs/podcasts than movies/tv-series. However, this did not seem to be the case. The students spent on average 3.94 hours a week on movie/tv-series with Norwegian subtitles, and on average 2.56 hour a week on movies/tv-series without Norwegian subtitles. Lastly, they only spent 0.60 audiobooks/radio programs/podcasts. There is no way to certainly measure if the *self-reporting bias* played a role in the reporting of the students, but they spent a considerably lower amount of time on the activities that might generally be viewed as more positive.

## 5.3 RQ 3: *Does total time spent on Extramural English improve the lexical richness of students?*

To look at this research question, I created the mentioned groups that can be seen in Table 6. It is interesting that these groups were so evenly divided and had almost the exact same number of participants in each group. The mean total time the students spent on the activities were *16.21* hours a week. This is comparable to a study from Sundqvist (2009), from a similar age group where the students were 15-16 years of age. In this study the mean time students spend on Extramural English in total was found to be *18.4* hours a week (Sundqvist, 2009, p. 116). As can be seen on Table 6, total time spent on Extramural English was the part of the results that showed the biggest difference in lexical richness scores. However, there was not found to be any correlation between the activities and lexical richness gain. In Sundqvist (2009) there was found a clear correlation between the time the students spent on Extramural English activities and their vocabulary scores. In my data analysis this was not the case. Several reasons have been mentioned for why this could be the case, such as age of participants, the tasks, the method of analysis, the cutting of texts etc. Furthermore, when combining all the activities another factor could be the scores I created for each student. I created groups based on the time they spent on each activity 1-4 hours became a 2.5 score, 5-10 hours became a 7.5 score and so on. When combining all the activities this could create a greater error margin because some students could be consistently on the higher side of the averages, or on the lower side. This questions the reliability of the findings. However, this was necessary due to the nature of the questionnaire, which was used for the TRAWL data collection, because it did not ask them to report exactly how much time they spent.

For further research it would be more accurate to find out exactly how much time the students spend on each activity. Sundqvist (2009) found that there was a difference between how effective the different Extramural English activities were in terms of vocabulary gain. This is an interesting area for further research concerning how this paper found little difference in scores between individual activities and total time spent on Extramural English. As mentioned previously in the theory section Dasic's paper from (2019) found non-significant differences between the gaming habits of students and their lexical richness scores. In this paper, I have used a larger sampling size, and still found non-significant differences, not only with regards to gaming but also the other activities. Auensen's (2019) paper also mentions that a larger sampling size would be interesting (Auensen, 2019, p. 51). The research focus of

that paper was different from mine, however, it also focused on lexical richness scores as a measurement of vocabulary.

## 5.4 Limitations and other possible factors

The first major limitation that should be mentioned is the sampling size. This research was conducted on 77-80 participants, depending on the activity in question. This type of study, which is based on surveys and text analysis, could be done with a much greater number of participants, and would yield more reliable results. Furthermore, all of the texts were gathered from a single day of writing in the 8th grade, which arguably, would not show a full picture of the lexical richness of students. Some could have had a bad day or gotten a topic they are not as familiar with. As mentioned in Laufer & Nation (1995) familiarity with a subject could affect the vocabulary scores of the students. A larger number of texts from the same participants could be a good addition to this type of research. Moreover, many of the answers were from different tasks, which could affect lexical richness. One could likely gather more accurate results if one was to design similar tasks for all participants. Lastly, the tasks could be designed in a way that all participants had to write exactly 250 words, which would eliminate the need to cut the texts, this could also help generate more accurate results.

The primary focus of this paper has been on Extramural English activities effect on lexical richness. However, since I had available data for other factors, such as total number of words, gender and off-list words, I looked at these as well in the results section. Firstly, with regard to total number of words, there was not shown to be a difference between total time spent on Extramural English, and the total number of words the students wrote. As shown on Table 6, the low group had more words than the moderate group and the high group had the most words. Even though I cut the texts to 250 words total number of words, this is something that could affect the lexical richness. One of the reasons for this is that the students who write so little that they have time to answer more tasks, might get different lexical scores. The shortest text analyzed contained 279 words, while the longest text contained 2154 words before being cut. This means that the scores from the shortest text and the longest text were quite far away from the mean number of words from all of the texts combined. Cutting the texts allowed for the metrics to be generally more reliable, however, some students that had written much more than 250 words might be negatively impacted. It should also be mentioned that these answers

were from 5-hour end of term tests, and the students who had written around 250 words, would have written very little in this time.

The gender division is also notable, and the 6-hour difference in mean total time spent per week between the two genders was substantial. Activities such as reading English on the internet, and playing computer games in English, were particularly different between the genders. The boys spent more time on these. From the participant group most of the students that spent the most time were boys. However, regarding the data analyzed it seemed that this time spent did not affect the lexical richness scores to any meaningful extent. An area of research that would be interesting to explore is what types of content the different gender engages with in these activities, and if there are major differences. As can be seen from the scores, there is clearly a difference in the amount of time that they spend. This is interesting, but it would also be beneficial to look at specifically the type of input the genders engage with and if it is at a similar level and contains similar content. Lastly, the off-list word would be an interesting topic for further research. As previously mentioned, many of the off-list words were spelling mistakes, and many of them were related to the task. It would be an interesting area of focus to see particularly what types of words the students use here, and how they learnt the different words. The input the students receive could affect both the more advanced (NGSL 3 and NAWL) and the off-list words that they use.

# 6.0 conclusion

In this paper there was not found to be any significant correlation between the time these students spend on Extramural English activities and their lexical richness scores. None of the metrics applied, such as, type-token ratio, flemma-token ratio, lexical density, lexical sophistication and lexical sophistication with percentages, had notable differences with regard to the time the pupils spent on the activities. Other findings were also discussed, such as gender differences between time spent on activities, the total words the students had written and the off-list words they used. The differences between the input types the students are engaging in has also been discussed in this paper, and this is an interesting area for further research. The research field of Extramural English is a growing one and methods for teachers to utilize this for their students learning could be very significant. Vocabulary is an aspect that Norwegian students have to focus on for their English learning (Norwegian Directorate for Education and Training, 2019, p. 2). Even though this study did not find significant

correlation between lexical richness and these activities, other studies mentioned such as Arndt & Woore, (2018), Aschcroft et al., (2018), Feng & Webb, (2020), Peters et al., (2016), Putman & Kingsley, (2009), Sundqvist (2009), found a positive correlation between some of these activities and the vocabulary gain of the participants. Factors such as age, input, participants, groupings, tasks, and method of vocabulary assessment have been mentioned as to why this was the case in these studies, but not in this paper.

In this study very few of the students engaged with audiobooks/radio programs/ podcasts, which were looked at individually. When looking at all the Extramural activities it was clear that other activities also had the issue that the students spent low amounts of time with these, this includes time spent reading English from books/newspapers/magazines, chatting, writing emails/sms in English, talking to someone in English, and other specified uses of English. With all of these activities the students spent below 1 hour on average. Therefore, the findings are hard to trust with regards to these particular activities. While it might be the case that the students do not gain as much vocabulary from these activities, it could also be that they do not engage with them enough. This would be an interesting area for further research, where more of the students were engaged with these activities. Several things have been mentioned with regard to why the activities did not yield results, such as, text cutting, tasks, the group sizing and so on. Similar studies with more resources could be conducted, with control factors related to the aforementioned factors that effect this study's reliability. Lastly, vocabulary is an important aspect of what it means to know a language. Digital arenas are widely available for almost all Norwegian 8th grade students. Exploring how they can be a tool for language learning is important, and figuring out how to best utilize these multimedia tools is essential for future teachers. Therefore, more research on how Extramural English effects students' vocabulary is crucial.

# 7.0 References

Arndt, H., & Woore, R. (2018). Vocabulary learning from watching Youtube videos and reading blog posts. *Language Learning and Technology*, *22*(3), 124–142. https://ora.ox.ac.uk/objects/uuid:e5159356-78e6-4560-bd94-8d7242b256f8

Ashcroft, R. J., Garner, J., & Hadingham, O. (2018). Incidental vocabulary learning through watching movies. *Australian Journal of Applied Linguistics*, *1*(3), 135-147. https://doi.org/10.29140/ajal.v1n3.89

Auensen, M. (2019). *The Correlation Between Lexical Richness and Norwegian Lower Secondary School EFL Teachers' Assessment of Written Compositions.* [Master's thesis, University of Agder]. Norwegian Open Research Archives. https://uia.brage.unit.no/uia-xmlui/handle/11250/2617582

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for. *Vocabulary learning and Instruction*, *3*(2), 1-10.

Browne, C., Culligan, B., & Phillips, J. (2013). *New Academic Word List 1.0.* https://www.newgeneralservicelist.com/new-general-service-list-1

Cobb, T. (2023). *Compleat VP v2.6 [computer program] Retrieved 28 february 2023 from https://www.lextutor.ca/vp/comp/*.

Creswell, J. W., & Guetterman, T. C. (2021). *Educational Research Planning, Conducting, and Evaluating Quantitative and Qualitative Research* (6th ed.). Pearson Education Limited.

Dasic, J. (2019). *Acquiring English Vocabulary Through Virtual Worlds : Exploring the connection between Norwegian EFL lower secondary pupils' gaming habits, essay grades and written lexical richness in light of their attitudes.* [Master's thesis, University of Agder]. Norwegian Open Research Archives. https://uia.brage.unit.no/uia-xmlui/handle/11250/2617514

Dirdal, H., Drange, E.-M. D., Evang, K. H. S. Ø., Garshol, L., Hamann, V. R., Hasund, I. K., Johansen, S. H., Vold, E. T., & Weisenberg, M. W. (2021). *TRAWL: Tracking Written Learner Language.* https://www.hf.uio.no/ilos/english/research/groups/trawl-tracking-written-learner-language/.

Dirdal, H., Hasund, I. K., Danbolt Drange, E.-M., Thue Vold, E., & Berg, E. M. (2022). Design and construction of the Tracking Written Learner Language (TRAWL) Corpus: A longitudinal and multilingual young learner corpus. *Nordic Journal of Language Teaching and Learning*, *10*(2), 115-135. https://doi.org/10.46364/njltl.v10i2.1005

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*(2), 245-260. https://doi.org/10.1023/a:1019637632584

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing. *Studies in Second Language Acquisition*, *42*(3), 499-523. https://doi.org/10.1017/s0272263119000494

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* SAGE.

Granger, S., & Wynne, M. (2000). Optimising measures of lexical variation in EFL learner corpora. In *Corpora galore* (pp. 249-257). Brill. https://doi.org/10.1163/9789004485211_020

Hasan, M., & Hoon, T. B. (2013). Podcast applications in language learning: A review of recent studies. *English language teaching*, *6*(2), 128-135.

Hasund, I. K., Drange, E.-M. D., & Hansen Torjusen, H. M. (2023). Tracking Written Learner Language. *Nordic Journal of Language Teaching and Learning*, *10*(2), I-XVIII. https://doi.org/10.46364/njltl.v10i2.1131

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *21*(2), 181-193. https://doi.org/10.1017/s0272263199002028

Indahsari, D. (2020). Using podcast for EFL students in language learning. *JEES (Journal of English Educators Society)*, *5*(2), 103-108. https://doi.org/10.21070/jees.v5i2.767

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning*, *63*, 87-106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, *53*, 61–79.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens. *Vocabulary Learning and Instruction*, *1*(1), 60-69.

Krashen, S. (1982). Principles and practice in second language acquisition.

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly, 49*(4), 757-786. https://doi.org/10.1002/tesq.194

Laufer, B. (1989). 25 What Percentage of Text-Lexis is Essential for Comprehension? *Special language: From humans thinking to thinking machines*, 316-323.

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

McCarthy, M. (1990). *Vocabulary*. Oxford University Press.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392. https://doi.org/10.3758/brm.42.2.381

McLean, S. (2018). Evidence for the Adoption of the Flemma as an Appropriate Word Counting Unit. *Applied linguistics*, *39*(6), 823-845. https://doi.org/10.1093/applin/amw050

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University press.

Nation, P., & Kyongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*(1), 35-41. https://doi.org/10.1016/0346-251x(94)00050-g

Norwegian Directorate for Education and Training. (2019). *Curriculum in English (ENG01-04)*. Established as regulations. The National curriculum for the Knowledge Promotion 2020. https://data.udir.no/kl06/v201906/laereplaner-lk20/ENG01-04.pdf?lang=eng

Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System*, *63*, 134-148. https://doi.org/10.1016/j.system.2016.10.002

Pinchbeck, G. G. (2017). *Vocabulary use in academic-track high-school English literature diploma exam essay writing and its relationship to academic achievement.* [Doctoral dissertation, University of Calgary]. Proquest Dissertations & Theses Global. https://ucalgary.primo.exlibrisgroup.com/permalink/01UCALG_INST/15o3ob6/cdi_proquest_journals_1923083867

Putman, S. M., & Kingsley, T. (2009). The Atoms Family: Using Podcasts to Enhance the Development of Science Vocabulary. *The Reading Teacher*, *63*(2), 100-108. https://doi.org/10.1598/rt.63.2.1

R Core Team. (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Read, J. (2000). *Assessing vocabulary*. Cambridge university press.

Richards, B. (1987). Type/Token Ratios: what do they really tell us? *Journal of Child Language*, *14*(2), 201-209. https://doi.org/10.1017/s0305000900012885

Rodgers, M. P. (2016). Extensive listening and viewing: The benefits of audiobooks and television. *The European Journal of Applied Linguistics and TEFL*, *5*(2), 43-57.

Schmidt, R. (1992). Awareness and Second Language Acquisition. *Annual Review of Applied Linguistics, 13*, 206-226. https://doi.org/10.1017/s0267190500002476

Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics*, *41*(4), 601-606.

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary* [Doctoral dissertation, Karlstad University]. Digitala Vetenskapliga Arkivet.

Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in teaching and learning: From theory and research to practice*. Springer.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323-352. https://doi.org/10.1023/a:1001749303137

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65-83. https://doi.org/10.1177/026553220001700103

West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longman.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

# 8.0 Appendix

*Appendix 1: TRAWL questionnaire.*

## SPØRRESKJEMA ENGELSK OG FREMMEDSPRÅK

| Etternavn: | Fornavn: | Landet du ble født i: |
|---|---|---|
| Ditt morsmål: | Morsmålet til forelder/foresatt 1: | Morsmålet til forelder/foresatt 2: |

Har du vært med i ESIT/TRAWL tidligere? ☐ ja  ☐ nei  ☐ vet ikke/husker ikke

Alder: _____  Kjønn: ☐ jente/kvinne  ☐ gutt/mann  ☐ annen kjønnsidentitet

Har du bodd i andre land enn Norge? Hvilke og når? _____
_____ (Fortsett på baksiden om nødvendig.)

Hvilke språk kan du i tillegg til morsmålet ditt? (Dette kan være språk du snakker hjemme eller med slektninger, språk du har lært på skolen eller språk du har lært på andre måter. Kryss av for om du kan lese, skrive, snakke eller forstå språket når du hører det.)

| Språk | Lese | Skrive | Snakke | Forstå |
|---|---|---|---|---|
| | | | | |
| | | | | |

Har du gått på skoler med et annet undervisningsspråk enn norsk (dvs. det språket læreren brukte i de fleste fagene)? Oppgi i så fall klassetrinn og språk nedenfor:

| Klassetrinn | Undervisningsspråk |
|---|---|
| | |
| | |

I hvilken klasse begynte du med **engelskundervisning?** _____

Har du bodd i et **engelsktalende** land? Kryss av for perioden.
☐ Nei, eller mindre enn 2 uker  ☐ 4-6 måneder
☐ 2-4 uker  ☐ 7-12 måneder
☐ 1-3 måneder  ☐ Mer enn ett år

| Kryss av for omtrent hvor mange **timer i uka** (utenom skole og lekser) du bruker til å... | Over 10 timer | 5-10 timer | 1-4 timer | Opp mot 1 time | Ingen |
|---|---|---|---|---|---|
| Lese **engelsk** på internett | | | | | |
| Lese **engelske** bøker/aviser/blader | | | | | |
| Spille dataspill hvor du bruker **engelsk** | | | | | |
| Chatte/skrive e-post/SMS på **engelsk** | | | | | |
| Samtale muntlig med noen på **engelsk** | | | | | |
| Se serier/filmer med **engelsk** tale og norsk teksting | | | | | |
| Se serier/filmer med **engelsk** tale uten norsk teksting | | | | | |
| Høre på lydbøker/radioprogrammer/podcast e.l. med **engelsk** tale | | | | | |
| Annet (spesifiser) | | | | | |