SPLLÄR 2023-018

GÖTEBORGS UNIVERSITET

Word Frequencies and Vocabulary Threshold Levels in EFL Textbooks

A Corpus Study

Samantha Read
Ämneslärarprogrammet

Degree essay: 15hp
Course: LGEN2A
Level: Advanced level
Term/year: VT/2023
Supervisor: Monika Mondor
Examiner: Christina Lindqvist

Keywords: Vocabulary, corpus, coursebooks, upper secondary, word frequencies, threshold levels.

Abstract

A significant number of studies have identified inadequate vocabulary exposure in EFL coursebooks as a problem. However, few studies have been carried out in the Swedish context. This observation, along with how the most recent revisions of the Swedish syllabi mention vocabulary explicitly, inspired this investigation of the vocabulary exposure in upper secondary coursebooks in the Swedish context. In total, text excerpts from four books were analysed through *AntWordProfiler* (Anthony, 2022), also revealing low exposure to high- and low-frequency words. However, since word frequencies are based on native speaker production rather than learner-produced language, their role as an accurate gauge for potential vocabulary exposure can be questioned. Thus, from the coursebook corpora a small number of words were selected and compared to the threshold division made by CEFR (Council of Europe, 2020) in the *English Vocabulary Profile* (Cambridge University Press, 2015c). Revealing a vast representation of words from all threshold levels in the majority of the coursebooks, the author suggests that future studies focus on analysing the percentage of the presence of different threshold levels instead of word frequencies in EFL coursebooks. Furthermore, it is strongly advised not to assign all vocabulary gains to a coursebook, since the results also indicate that the imbalance between content and vocabulary exposure will likely remain during the compilation of coursebooks.

Table of Contents

1 Introduction

Coursebooks will always exist in education, regardless of the level or subject. Even though their use is optional for teachers, at least in the Swedish context, coursebooks are there to give structure and support since they are designed to reflect both curricula and syllabi (O'Keeffe, 2013, p. 1; Hutchinson & Torres, 1994, p. 317). Notwithstanding, in the English Language Teaching (ELT) context, the use and quality of coursebooks have since long been a subject of debate due to their inability of keeping up with recent updates of the syllabi and curricula (Hutchinson & Torres, 1994, p. 316; Sheldon, 1988). Sheldon argues that instead of disregarding the coursebooks when they fail to live up to singular updates, they should be evaluated for their ability to support the fundamental and lifelong aspects of learning a language (Sheldon, 1988, p. 245).

The belief that a person's ability to use a language is determined by their vocabulary knowledge is widely accepted in the research community (Schmitt, 2008, p. 329; Cook, 2016, p. 58; Brown & Lee, 2015, p. 480). Within the research field of reading comprehension, vocabulary knowledge is a frequently used gauge (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2013; Schmitt, 2008). Despite this, reading has had an integral part of the syllabus of the Swedish upper secondary for a long time, whereas vocabulary has not been explicitly articulated in the syllabus until recently (Skolverket, 2020).

When incidentally acquiring vocabulary through reading, meaning without explicit focus on learning, the retention of words depends on the frequency of encounters with a word (Nation, 2015, p. 136, Ellis, 2002, pp. 150-152). Word frequencies can be further divided into two types: appearance in the language in general, or appearance within a specific text, also called repetition. The former type, how frequent or infrequent words are in language use, is commonly used as a gauge when calculating how many word families need to be known to sufficiently understand a text when reading, also called lexical coverage, referring to a knowledge of 95-98% of a text (Laufer, 1989, p. 21; Nation, 2006, pp. 71-72). The latter type, word repetition, measures which words score the highest in terms of recycling, in other words, which words are more likely to be retained from reading a text. These two aspects of frequency are pertinent when it comes to learning vocabulary incidentally through reading (Bergström, Norberg & Nordström, 2022, pp. 2-4; Yang & Coxhead, 2020, p. 598; Matsuoka & Hirsh, 2010, p. 57-58; Sun & Dang, 2020, p. 3, among others). On the one hand, if the vocabulary needed to obtain a 95-98% coverage is far above the knowledge of the student, or if the rate of repetition is too low, the text will be too difficult. On the other hand, if the coverage is equal to, or lower

than the vocabulary knowledge of the student, but the repetition rate is low, chances of vocabulary gains are small. Thus, we can deduce that if English as a Foreign Language (EFL) students read texts with insufficient exposure and repetition to certain words their vocabulary development will suffer.

The problem of inconsistent vocabulary exposure through EFL coursebooks previously identified in an international context (Alcaraz, 2009; O'Loughlin, 2012; Matsuoka & Hirsh, 2010), hypothesized the problem pertaining to coursebook designers not prioritizing vocabulary (Alcaraz, 2009, p. 71). A number of studies in the Swedish context addressing the vocabulary input of EFL coursebooks designated for the primary and early secondary school years found the same inconsistency (Norberg & Nordlund 2018; Nordlund & Norberg, 2020; Bergström, Norberg & Nordlund, 2022). The aforementioned hypothesis of the origins of the problem was confirmed by Bergström, Norberg, Nordlund (2023, p. 165) who found that coursebook designers compose materials based on intuition rather than recent EFL vocabulary research.

Interestingly, although vocabulary frequencies are mainly based on native speaker (NS) language found in corpora, Skolverket has based their course levels for English 5-6 on the threshold levels B1-B2 described in the CEFR (Skolverket, 2022), which in turn are not based on NS language, but on a corpus of learner language (Cambridge University Press, 2015a). With reference to this, it is relevant to question the previous use of vocabulary frequencies as a gauge for the complexity of texts in EFL coursebooks in the Swedish context. The absence of studies regarding this matter as well as coursebook studies regarding the upper secondary years is thus particularly striking.

As suggested above, a varying degree of exposure to word frequencies due to intuition-based coursebook compilation (Bergström, Norberg, Nordlund (2023, p. 165) seems to monopolize the EFL coursebook research not only nationally, but also internationally. Simultaneously, the possible incompatibility between word frequencies and threshold levels may be further distorting an accurate way of addressing the issue of the lack of systematic vocabulary exposure in said coursebooks. In recognition of these gaps, the aim of this study is thus, twofold: to investigate the potential vocabulary gains made from EFL coursebooks, as determined by the word frequency representation in the text excerpts, and to question the use of word frequencies as a gauge for potential vocabulary gains. The following research questions will guide the study:

1. What is the representation of the first 3,000 word frequencies in all EFL coursebooks, and to which extent are words from these repeated?

2. How does the representation of word frequency bands progress, and to which extent are their words repeated?
3. Is there a correspondence between the appearance of the words in the frequency bands and the categorisation of words into threshold levels made by the *English Vocabulary Profile* (Cambridge University Press, 2015c)?

Before operationalising the research questions for this study in detail, the means by which we can define and measure vocabulary goals for EFL need to be presented (2), followed by a literature review of relevant studies (3). Subsequently, the choices behind the material accounted for in the method section will be elucidated (4), followed by the answering, and limitations of the research questions in chronological order in the results (5). Lastly, the results will be summarised, followed by a discussion of their potential pedagogical implications (6-7).

2 How to Define and Measure Vocabulary Goals

First, in this section a description of how vocabulary goals are defined through lexical coverage and threshold will be made, followed by a brief summary of key studies in the field. Second, how the vocabulary skill is expressed in the educational context will be illustrated, as well as how this translates to word frequencies and threshold levels. This section will close by accounting for common tools used to measure vocabulary levels.

2.1 Defining Vocabulary Goals Through Lexical Coverage and Threshold

Since high exposure to various word types is essential for vocabulary acquisition through reading, it is important to understand how recommendations have been previously expressed. According to Nation (2013, pp. 11-14), the parameters are defined by a) the size of the language, b) the vocabulary knowledge of a native speaker (NS), and c) an estimation of how many words are needed when reading different text types.

Addressing the size, the manner in which words will be calculated and categorised needs to be decided first. The Oxford English Dictionary (OED) has 600,000 entries (OED, 2023), each containing all words belonging to the same word family, for example, *talk, talkative, talking*, etc. By learning the base word, *talk,* for example, the learning of the rest of the words in this word family is likely to happen automatically, depending on the level of the learner

(Goulden et al., 1990, p. 344; Nation, 2013, p. 11). Consequently, to estimate the size of a language as a base for determining vocabulary-gaining goals, the preferable unit use is word family, thusly, landing on a number of around 70,000-word families (Nation, 2013, p. 12). Said word families are in corpus studies divided into word frequency bands or lists, which in some corpora occasionally have been further divided into lemmas (Leech et al., 2001), for example in the British National Corpus (BNC, n.d.). Lemmas are similar to word families but with a more restricted word division. A word family contains all derived forms related to a headword regardless of its area of use, whereas a lemma contains the inflected and reduced forms of a headword as long as they belong to the same part of speech. For example, a lemma divides the noun and verb form of a word, such as *walk* and *walk,* into different lemmas, whereas a word family does not make this division (Nation, 2013, p. 10; Cobb, 2017).

Calculations of the vocabulary knowledge of NS after the recent technological burgeoning seems to be prominently absent, which is one of the many reasons why their modern-day validity is questioned. The most recent studies found, are from Goulden et al. (1990, p. 356) and Zechmeister et al. (1995, p. 210), who estimated that the average NS knows 17,000 word families. Of the 183 participants ranging from college students to older adults, 163 were from the U.S. (Zechmeister et al., 1995, p. 203) and 20 were of unspecified NS nationality (Goulden et al., 1990, p. 356). Nation (2013b, p. 13) not only claims that 17,000 is a low estimation since no proper nouns were included but simultaneously also challenges its accuracy since variation between individual NS can be considerably high. Adding to this, even in the inner circle countries where English is a native language, Australia, Canada, New Zealand, USA, Great Britain, and Ireland, language use can differ in terms of vocabulary, etc., which raises the question of which NS should be the goal (Cook, 2016, p. 179). Be that as it may, the supposed NS goal, regardless of both nationality and level of education of the speaker, has been repeatedly used as a reference goal in studies determining the number of words necessary to read texts of various genres, thus, making its presence in this paper necessary.

Reading when not understanding is both frustrating and demotivating, but how does one reach the state of comfortable reading comprehension? When pausing to look up a word, it disturbs the reading experience (Hu & Nation, 2000, p. 403, Laufer, 2013, p. 871). Therefore, limits as to how many unknown words should appear in the running text have been set based on assumptions shown in Hu and Nation (2000, p. 405), but still lack clear consensus (Laufer & Ravenhorst-Kalovski, 2013, p. 16). For a lexical coverage of 95% of a text, (referred to as vocabulary load by Yang & Coxhead, 2020, p. 599) one word in 20 will be unknown, and for a text coverage of 98% one word in 50 will be unknown (Hu & Nation, p. 405).

Furthermore, controversy regarding which of these two coverage levels should be recommended has been actively ongoing in the past. Laufer (1989) suggested that 95% coverage is sufficient for comprehension of texts of a 'general academic nature' (p. 321) reflecting a knowledge of 5,000 words, without Laufer referring to word families. In a similar vein, Laufer and Ravenhorst-Kalovski (2013, p. 26) suggested two thresholds, one of 98% and one of 95% comprehension, on the basis that the latter can give adequate room for further vocabulary gains to be mastered within a course. Nevertheless, when Hirsh and Nation (1992) found that when reading youth novels, a knowledge of the first 2000-word frequency bands was not enough to reach a coverage of more than 89-90%, a knowledge of 5,000–7,000-word families was therefore suggested, reaching a coverage level of 98%. Adding to advocates of the 98% coverage, were Hu and Nation (2000, p. 422) and Nation (2006) who also specified the target number of word families of 8,000-9,000 when reading newspapers and novels (pp. 71-72).

Comparing the two standpoints of 98% and 95%, the former represents the goal for vocabulary knowledge, whereas the latter, prefers the term 'threshold' since it denotes the minimum requirement (Ravenhorst-Kalovski, 2013, p. 15). Thus, the standpoint which advocates the 95% coverage level gives more room for the acquisition of vocabulary levels of EFL students to take place. Accordingly, a 'threshold of 95%' is a more reasonable goal when it comes to EFL coursebooks, since their texts come accompanied with vocabulary lists and exercises, as expected in the educational context. Adding to the dynamic character of language learning, van Ek (1976), whose research laid the foundations of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2020), described "foreign language ability as *skill* rather than *knowledge*" (p. 5).

2.2 Defining the Vocabulary Skill in the Educational Context

In the composition of the syllabus for the courses of English step 5-7 in the Swedish context no reference to neither Nation's 98% coverage level nor Laufer's lexical threshold have been made, therefore their applicability in the Swedish context can be questioned. Instead, Skolverket have based their levels of general language proficiency requirement for courses 5-7 on the levels B1.2-B2.1 (Skolverket, 2022, p. 7), from the scale A1-C2 in the CEFR (Council of Europe, 2020, p. 36). Worth pointing out is that the maximum level, C2, is not based on the "performance of an idealised 'native speaker'" but rather what is "intended is to characterise the degree of precision, appropriateness and ease with the language which typifies the speech

of those who have been highly successful learners" (Council of Europe, 2020, p. 36). In this context, the requirement levels B1.2-B2.1 are in the CEFR (Council of Europe, 2020, p. 131) labelled as "Independent User" (p. 36), and in terms of vocabulary range expressed as follows.

Description of vocabulary range for threshold level B1-B2 (Council of Europe, 2020, p. 131).

B2	Can understand and use the main technical terminology of their field, when discussing their area of specialisation with other specialists. Has a good range of vocabulary for matters connected to their field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. Can produce appropriate collocations of many words/signs in most contexts fairly systematically. Can understand and use much of the specialist vocabulary of their field but has problems with specialist terminology outside it.
B1	Has a good range of vocabulary related to familiar topics and everyday situations. Has sufficient vocabulary to express themselves with some circumlocutions on most topics pertinent to their everyday life such as family, hobbies and interests, work, travel and current events.

Attempts have been made regarding the translation of these vocabulary levels into the required knowledge of word families by Nation (Victoria University of Wellington, n.d.A). In a document titled *Vocabulary and the CEFR* a table is illustrated, without reference to calculation, showing that the vocabulary knowledge of students within the B1-B2 levels, corresponds to 2,000- 4,000 word families, plus 1000-2000 word families of specialised vocabulary (Victoria University of Wellington, n.d.A), meaning academic and technical vocabulary (Nation, 2013, p. 19). Because of the unsure rationale behind these numbers, one can question how, and whether if at all, such numbers can be used as a goal or not.

Since the general CEFR levels refer not only to English, but also other languages, the CEFR has together with Cambridge University Press also developed the *English Profile* (2015b), which further describes each proficiency level based on a corpus consisting of learner-produced language from students all over Europe (Cambridge University Press, 2015a). Found in the *English Profile* (2015b), is the *Vocabulary Profile Online* (Cambridge University Press, 2015c), which makes no specific word family indications, but divides words into previously described levels from the *English Profile* (Cambridge University Press, 2015a). Further specifications of this online resource are found in the method section (4).

2.3 Corpora and Software Commonly Used in Linguistic Studies

When researching the characteristics of language or coursebooks, a corpus, defined as a large "collection of written or spoken material stored on a computer and used to find out how

language is used" (Cambridge Dictionary, 2023) is the most apt tool for reference (McKay, p. 126). Because the definition of *corpus* is so wide, it can also represent a collection of digitally stored coursebooks. Consequently, the material used for a coursebook analysis therefore often consists of two corpora, one representing the target coursebooks, and one corpus consisting of a vast collection of written or spoken language representing genres such as newspapers, novels, magazines, movies, TV, and the likes. The most used corpora are the Corpus of Contemporary American English (COCA), and the BNC (n.d.).

The COCA was created in 1990 and includes over one billion *tokens,* which differ in meaning from *type* in the way that the latter represents a word, and the former represents the frequency of word appearances, regardless of whether they are the same word or not. While the BNC, created in 1991 (University of Oxford, 2022), is a corpus of significantly smaller size, 100 million words, its contents are not genre-balanced as that of COCA, which has all text and speech divided into the genre from which they originate, such as letters, newspapers, TV, etc. Other significant differences between the two corpora are that the BNC represents British English and has not been updated since 2007 (University of Oxford, 2022), whereas COCA represents American English and was updated as recently as 2019 (Word Frequency Data, n.d.). Since corpora consist of language gathered during a limited period of time it is not guaranteed to contain all 600,000 entries of the OED (2023) and therefore cannot reflect the English language in its entirety.

Taking the comprehensive size of corpora into account in combination with specific purposes of linguistic studies, the analysis is often further facilitated using programs such as *Range* or *AntWordProfiler*, to mention a few (Victoria University of Wellington, n.d.B). These have built-in functions that can directly carry out type-token ratio (TTR) and word frequency band calculations. Nevertheless, the size of corpora impedes the performance of the programs, hence, lists representing base words of word families divided into frequency of appearance are used. In the case of the list based on the BNC/COCA (Anthony, n.d.), its first frequency band represents the 1000 most frequent base words in English, and the second frequency band represents 1000 slightly less frequent base words, etcetera. The first 3,000 word families are considered to be high-frequency, then from 3,001 to 8,000 is considered mid-frequency, and 8,001 and up represents low-frequency (Nation, 2013, p. 18). Apart from being based on BNC and COCA, the lists used in the linguistic software described above can also be based on the 2284 word families considered to represent the language of highest learning value for EFL students (Smith, 2023), known as the General Service List (GSL), although these are not frequency-based (Nation, 2013, p. 18). An additional list, called the Academic Word List

(AWL), contains 570 word families of academic language not represented in the GSL (Coxhead, 2000, p. 222).

When it comes to determining which words are more necessary to learn in the Swedish context, no attempts have been made since Thorén (1976, p. xii), who mapped the learning order of 9,600 from the mid- elementary to the end of the upper secondary school years. In its making, several factors were taken into consideration, such as cognates in common, prioritising Swedish words that have a higher word frequency than in English, for example, *midsummer,* and removing words assumed to be learned in the initial years in the English classroom, such as months and the days of the week (Thorén, 1976, p. 16, p. 18). Furthermore, since the compilation of this list took place before the technical evolution, input was limited and thus, both the use and necessities of the EFL students were presumably easier to map than they are today. Owing to its limited use of only being valuable in the Swedish context, no one has so far updated and used this in corpus research.

As previously mentioned, earlier calculations have shown that the English language can be divided into approximately 70,000 word families, yet the lists available for download rarely encompass more than 34,000 base word lists (Anthony, n.d.). Still, this represents twice what the average NS knows (Zechmeister, et al., 1995, p. 210), equalling a wide representation of language. Necessary to remember is that since these corpora are based on NS language the goals may clash with the ones of CEFR which are based on native speaker knowledge (Council of Europe, 2020), or as Sun and Dang (2020) put it:

> Considering corpus-based information in relation to the vocabulary knowledge of students who actually use these textbooks would provide better insights into the vocabulary load of the textbooks for their users. (p. 1).

In light of the theoretical background, investigations of vocabulary coverage and progression in EFL coursebooks likely need to compare and re-evaluate previously established vocabulary goals which up until now have guided research. If these are not compared to each other, there is a risk of incompatibility and thusly, epistemic failure.

3 Previous Research: Evaluating Word Frequencies in Coursebooks

The use of coursebooks has been and continues to be both praised and discredited. As previously mentioned, on the one hand, they can be timesaving and give structure to the course planning for teachers and on the other, coursebooks seem to fail to keep abreast of recent

educational research and school policy updates (Sheldon, 1988, p. 237; Hutchinson & Torres, 1994, pp. 316-317). As a result of delays in necessary adaptations, coursebooks may fail to comply with modern pedagogical needs. Hutchinson and Torres (1994) explain that therefore, "[s]tudent teachers are taught that good teachers do not follow the textbook but devise their own curriculum and materials." (p. 316). Simultaneously, it is common to be told by practising teachers that whenever the time is scarce and planning time suffers, the coursebook is indispensable. The dichotomous role of the coursebook makes it an interesting subject for research.

The field of EFL research mapping the lexical coverage of coursebooks can be divided into two types of investigations. First, studies measuring only the lexical coverage will be depicted. Second, since incidental vocabulary acquisition in lack of deliberate focus on vocabulary requires a great deal of repetition of words within a text to a higher degree, studies also including recycling of words will be described. Lastly, there will be a summary.

3.1 Lexical coverage

Throughout the world, the use of coursebooks in EFL varies in terms of publisher options and national school policies. While in some countries it is common to base the complete course content on a coursebook (O'Loughlin, 2012, p. 256), other countries might even have a set selection and progression of school material spanning over many years (Alsaif & Milton, 2012, p. 24), whereas the use of coursebooks is in some educational systems seen as complementary or even voluntary. In places where the selection of coursebooks is predetermined and read from beginning to end, one can assume that little space is left for additional reading through novels, articles, and other text genres. Furthermore, if the daily contact with English or the extracurricular reading habits of these students are low, coursebooks will be the primary source for vocabulary gains (Alsaif & Milton, 2012, p. 22). These parameters are decisive when it comes to interpreting the results of coursebook analyses.

As stated before, high-frequency words will be learned first, followed by the less frequent words. According to Schmitt and Schmitt (2012, p. 486), the first version of GSL launched in 1953, representing 2,284 word families (as referred to in Smith, 2023), has been the indestructible source of reference, in outlining the most important words to initially learn for EFL students. Hence, there is a conflict between if students should focus on learning the slightly outdated words of the GSL list or the first 3,000 most frequent word families in a general-purpose corpus (Schmitt & Schmitt, 2012, p. 498). Therefore, on the one hand, there is

a large number of coursebook studies which have investigated to which extent EFL coursebooks reach a coverage of these 2,284 word families (e.g. O'Loughlin, 2012, p. 263; Matsuoka & Hirsh, 2010, p. 61), and on the other, a substantial amount of research has used the coverage of the first 3000 word families as their point of departure (Yang & Coxhead, 2020, p. 603; Sun & Dang, 2020; Bergström, Norberg & Nordlund, 2022).

Using the program *Range* (Victoria University of Wellington, n.d.B), without specifying a list of base words, Alcaraz examined if vocabulary had been prioritised in a coursebook used in the third year of Spanish primary education (Alcaraz, 2009, p. 64). The representation of word frequencies in the book revealed that the first 2,000 word families represented 72% of the book as a whole, the first 1000 being 56 % but the second 1000 only 16%, a number which is low considering their importance for vocabulary gains. Furthermore, the uncovering of the remaining 18% of uncategorised words (Alcaraz, 2009, p. 66), being a relatively high number, could have further revealed the potential vocabulary gains from word families above the first two frequency bands, consisting of 1000 word families each.

A similar imbalance between the coverage of the first and second word frequency bands was found by O'Loughlin (2012) who at a Japanese university mapped the word exposure in an EFL coursebook series designated for beginner to intermediate levels (p. 258). Using Cobb's *VocabProfiler version 3* (Cobb, 2009) which compares text content to the GSL list (O'Loughlin, 2012, p. 258), the calculations showed that the coverage of word families from the first and second word frequency bands were slightly imbalanced, namely 87,7% versus 55,8%, respectively (p. 262). However, O'Loughlin (2012) points out that many of the words in the GSL represent an obsolete use of language since words such as *plough* and *shilling* are present and words such as *email* and *online* are not (p. 263). Since a substantial amount of the word content of the GSL appears to be distant from modern language use, this may have caused the program to fail to recognise words of a more modern nature. Subsequently, these words might have been categorised as not part of the GSL, thus, distorting the true representation of the content of the word frequency bands.

In Saudi Arabia where the syllabus content and progression are predetermined, Alsaif and Milton (2012) used RANGE (Victoria University of Wellington, n.d.B), an unspecified reference list, and *LexTutor* (Cobb, n.d.) to investigate the vocabulary load of 22 coursebooks divided into course levels for ages 11-16 (Alsaif & Milton, 2012, p. 25). For the content of the complete coursebook set, only 80% of the first two frequency bands were represented, whereas up to the fifth frequency band, only 55% were represented (Alsaif & Milton, 2012, p. 32). Moreover, during a vocabulary size test on an unspecified number of students at an unspecified

point in time, scores showed that the students had learned on average 40% of the vocabulary from the coursebooks (Alsaif & Milton, 2012, p. 32). Adding the low learning rates to the low representation of word frequency bands equals limited vocabulary gains, especially when considering that all courses apart from between years 10-11, included complementary exercise books, giving students extra vocabulary practice (Alsaif & Milton, 2012, p. 32). However, given the unspecified nature of the vocabulary size test, these results need to be questioned.

3.2 Adding Word Recycling Frequencies

As stated in the introduction, pertinent to incidental vocabulary acquisition through reading is also the repetition of words within a text. The retention of words seems to be favoured by a repetition of at least 10 times within a text (Pellicer-Sanchez & Schmitt, 2010, p. 42; Webb, 2007, p. 60). Therefore, a great number of studies in the field of vocabulary recycling frequency in coursebooks will also be taken into consideration. The density by which words are repeated within a text affects vocabulary acquisition, since a more spaced repetition may aid the retention of words in long-term memory, according to Matsuoka and Hirsh (2010, p. 58). In their study they found that 90% of an upper-intermediate coursebook (Matsuoka & Hirsh, 2010, p. 58) contained language of the word families represented in the GSL (p.61), using Range (Victoria University of Wellington, n.d.B). Although the density remains unspecified, the repetition of 33% of words in the second frequency band appeared at least seven times, whereas 33% of some words only appeared once, whereby explicit vocabulary teaching is recommended by the authors (Matsuoka & Hirsh, 2010, p. 65).

Results from corpus studies of coursebooks indicating that a vocabulary of between 6,000-9,000 word families was needed to read coursebooks at the secondary school level in China were found by Sun and Dang (2020) and Yang and Coxhead (2020). Investigating coverage and rate of repetition in a corpus of 11 Chinese high school coursebooks, Sun and Dang (2020) found a regression in the coverage between the books (p. 7). Using *Range* (Victoria University of Wellington, n.d.B) and BNC/COCA lists, the representation of both coverage and repetition of the first three frequency bands and their words were found to reach the highest score in the last three books of the series (Sun & Dang, 2020, pp. 7-8). To reach a coverage of 98% of the written texts in the corpus, an average score showed that a vocabulary knowledge of 7,000 word families was needed.

Using a smaller corpus, but the same selection of software and corpus list, Yang and Coxhead (2020) analysed two upper-secondary coursebooks. A more linear coverage was

revealed between the coursebooks since the first book required 95% coverage at the first 3,000 word families and 98% at 5,000 word families, whereas the second book required 95% coverage at 5,000, and 98% at 6,000 word families, respectively (Yang & Coxhead, 2020, p. 603). Also analysing the development of the exposure to word families within the books, Yang and Coxhead (2020, p. 604) found that the first book showed progression in terms of higher requirements for a 98% coverage, whereas the second book shows fluctuating requirements for the same level coverage between units in the book. The asymmetrical results concerning coverage requirements found in Yang and Coxhead (2020) as well as in Sun and Dang (2020) may suggest that vocabulary development is prioritised to a higher degree in earlier stages of EFL courses than later. There may be several reasons for this, e.g. that vocabulary progression may after initial practice be expected to be taken over by students as they develop linguistic skills and independence, or it may be an arbitrary effect. Nevertheless, Nation (2013) emphasizes that often a coverage of 95% exceeds 4,000 word families (p. 26), meaning that acquisition of these needs to take place somehow, be it through tutoring or incidental learning through reading, as previously recommended by Matsuoka and Hirsh (2010, p. 65).

Next, we take a closer look at the studies in the Swedish context analysing coursebooks in a wider sense. Most of these studies have focused on the elementary to early secondary school years (Bergström, Norberg & Nordlund, 2022; Nordlund & Norberg, 2020; Norberg & Nordlund, 2018). Using *LexTutor* (Cobb, n.d.), Norberg and Nordlund (2018, p. 466) compared the vocabulary load of seven textbooks in middle school to the older version of the NGSL (Browne, 2013) and a corpus of language produced by NS children (Roessingh & Cobb, n.d., referred to in Norberg and Nordlund, 2018, p. 466**)**. Without specific data on the findings of the amount of high-frequency vocabulary nor the extent to which words were recycled, a number greater than anticipated was found pertaining to the less frequent vocabulary (Norberg & Nordlund, 2018, p. 469). Norberg and Nordlund (2018, p. 469) aptly conclude that even though high-frequency words are of high value in the principal years of EFL, constructing texts only containing the presence of such words would comprise strange and unnatural texts, just as a higher presence of lower-frequency words would in a text designated for later years of EFL.

Regarding the rate of repetition of words in middle school EFL coursebooks, Nordlund and Norberg (2020) investigated if a low rate of repetition would be compensated by the nature of vocabulary focus found in the exercises of the books. Although approximately 74% of the seven coursebooks investigated contained five or fewer repetitions of words (Nordlund & Norberg, 2020, p. 98), the majority of the vocabulary exercises focused on incidental acquisition (Nordlund & Norberg, 2020, p. 104, p. 107). Even though vocabulary exercises with

an incidental acquisition focus may seem contradictory, the authors' (Nordlund & Norberg, 2020) definition of these was "exercises [which] commonly do not have a structure that tries to draw students' attention to the language feature to be learned in an explicit way" (p.104). The lack of explicit focus on vocabulary as seen in the studies of the literature review so far, largely obstructs vocabulary gains.

Investigating the ratio between high- and low-frequency words, Bergström, Norberg and Nordlund (2022), using *LexTutor* (Cobb, n.d.) and a BNC/COCA list (Anthony, n.d.), but investigating five series of coursebooks used in school years 7-9 (Bergström, Norberg & Nordlund, 2022, pp. 6-7). The proportion of high- and mid-frequency word levels was 87.4-91.4% and 2.7-3.5%, respectively, throughout all books (Bergström, Norberg & Nordlund, 2022, p. 9), the recycling of which reached 93-94% for high-frequency and 0.5-1.4% for mid-frequency, based on a repetition criterion of 10 times or more (Bergström, Norberg & Nordlund, 2022, p. 12).

As previously mentioned, the presence of international research identifying the inconsistency of vocabulary exposure in EFL coursebooks from the elementary school level to the university level enhances the relevancy of the issue at hand. The main findings of the studies regarding coverage, present the absence of medium- to low-frequency words in the coursebooks as well as inconsistent repetition of target words. The studies which also scrutinise the recycling of vocabulary, found both unpredictability and sharp decrease regarding both rate of recycling and vocabulary progression. Wherefore, intuition rather than vocabulary research findings is likely practised by coursebook designers not only nationally (Bergström, Norberg & Nordlund, 2023, p. 165), but also internationally.

The literature review indicates the importance of investigating and comparing coverage, word repetition, and progression of difficulty level in coursebooks from different publishers for the upper secondary. Furthermore, to appropriately contextualise the linguistic needs of Swedish-speaking EFL students, the use of the list created by Thorén (1976) might have been a remedy. However, probably owing to its outdated use and the unavailability of an existing digital version, none of the studies in the Swedish context used it. An additional gap is created in the absence of studies comparing the correlation between word frequency bands and threshold levels. In recognition of these gaps, this study intends to shed some light on the vocabulary coverage of EFL coursebooks for the upper secondary, as well as initiate the scrutinization of the compatibility of word frequencies and threshold levels.

4 Method, Material and Analysis

To operationalise the research questions concerning the coverage of word frequency bands in EFL coursebooks of the Swedish upper secondary and to affirm or disprove the correspondence between word frequency divisions and the threshold levels used by the CEFR (Council of Europe, 2020), this section is dedicated to accounting for the method. First, the material selection process and the rationale behind the choices of coursebooks and their texts will be explained. Second, the benefits and limitations involved in the choices of software, website tools, and corpus list will be explained. Last, a description of how the coursebook corpus was created will follow, along with a brief summary of how the research questions will be answered.

4.1 Material Selection

All studies carried out in the field of word frequency analysis in coursebooks have their criteria for material selection depending on their educational context and specific research questions. In this specific study, regarding the Swedish upper secondary context, the main interests are to elucidate the exposure to word frequency bands in the texts selected for reading in the coursebooks and to confirm if previously established word frequency divisions correspond to the threshold levels used by the CEFR (Council of Europe, 2020, p. 36). Departing in these interests, below follows a justification of what books were chosen and why, followed by a delimitation regarding which texts were chosen within them, and a description of the categorisation of their genres.

4.1.1 Coursebook Selection

Before coursebook publishers were chosen, the desired criterion for selection was to find out which series of books were a) the most sold or printed, and b) of most recent edition. Thus, the goal was to include the most widely used and most recent textbooks in the Swedish upper secondary school context. However, since publishing information was not publicly available, the second option was to use a selection of convenience, i.e., by contacting publishers of EFL textbooks to see who would answer and share access to their books first. These were Liber, publisher of *Blueprint A-B* (Lundfall & Nyström, 2017; 2018), and Gleerups, publisher of *Viewpoints 1-2* (Gustafsson & Wivast, 2017; 2018).

Both coursebook series, *Blueprint* and *Viewpoints,* have published books for all three courses in English offered at the upper secondary school in Sweden, steps 5-7. However, to limit the scope and make the workload manageable, the focus of this study will be on the books

between courses 5-6. The underlying rationale is that since these two courses are compulsory, their coursebooks should make a true representation of what level students need to have reached when finishing upper secondary school.

4.1.2 Material Delimitation

To create an appropriate corpus of the chosen textbooks to analyse the vocabulary level of the texts in the books, the main interest was to only include texts selected by the authors, not written by them, with the aim of deciphering if vocabulary acquisition was a criterion for their text selection. Accordingly, no instructions were included nor were sections on grammar, exercises, or vocabulary lists since the words were already appearing in the texts themselves. Furthermore, sections focusing on listening were also excluded, unless they included a full script of the dialogue, which was the case for one text only. The exclusion of these texts was thought to avoid generating an overrepresentation of certain phrases or words which do not occur in everyday situations but tend to be so in school contexts, such as *write, present, discuss*, etc. Thus, the selected texts consist of excerpts of fiction and non-fiction only, divided into the following genres: fiction, news items, poems, informational texts, and other (see Figure 1).

The categorisation of the texts in *Blueprint A-B* (Lundfall & Nyström, 2017; 2018) was aided by the pre-categorisation made in their table of contents. The categories representing the lowest frequency of texts in *Blueprint A* (Lundfall & Nyström, 2017) were a timeline, sorted as *Informational Texts*, and a comic strip, song lyrics, a movie script, and a full-text script of a listening section, which were sorted as *Other.* The texts categorised as *Informational Texts* in *Blueprint B* (Lundfall & Nyström, 2018) were two factual texts and one report, and the ones categorised as *other* were two advertisements and one letter. Regarding *Viewpoint 1-2* (Gustafsson & Wivast, 2017; 2018), there was no table of contents, resulting in manual categorisation for all texts which might challenge its reliability. Categorised as *Other* in *Viewpoint 1* was a theoretical text and in *Viewpoint 2* song lyrics, and two essays. Categorised as *Informational Texts* in both books were biographies.

Figure 1. Text categorisation.

	Fiction	**Articles**	**Poems**	**Informational Texts**	**Other**	**Total**
Blueprint A	**6**	**10**	**3**	**1**	**4**	**24**
Blueprint B	**14**	**7**	**8**	**4**	**3**	**36**
Viewpoints 1	**7**	**4**	**2**	**1**	**1**	**15**
Viewpoints 2	**10**	**4**	**6**	**1**	**3**	**24**

The identified variation in the number of texts and their types in the books may have affected not only the size of the coursebooks but also the outcomes of which words are represented. If one book contains a wider selection of for example news articles, it might score a higher content of mid- or low-frequency vocabulary, since this is where specialised vocabulary is represented (Nation, 2013, p. 19). Therefore, the categorisation of texts is particularly pertinent when answering the research question in regard to comparing the progression of word frequency bands between all four books, as will be seen in section 5.3, page 26.

4.2 Corpora and Programs

Below follows an account of the choice of software, the website used for threshold comparison, and lists of word families from corpora. Finally, how the coursebook corpus was compiled is described. All four components have their possibilities and limitations regarding the answering of the research questions, which will be accounted for along the way.

4.2.1 Software and Website Tools

The software that will be used for word frequency and repetition analysis is *AntWordProfiler* (Anthony, 2022). In its default settings, the corpus lists GSL and AWL can be used to analyse the lexical coverage of the 2284 learner language word families (Smith, 2023), and an additional 570 word families of academic language (Coxhead, 2020, p. 222). However, since the former list is a slightly outdated version of English, and represents a low number of word families, a more rigorous list is required. Even though the use of the AWL would have been beneficial, the exclusion of one list is interpreted as the exclusion of both lists, rendering the use of any of the two default lists unusable. Moreover, a functional version of the AWL list was not to be found, thusly, the BNC/COCA list was downloaded from Lawrence Anthony's webpage (n.d.), the choice of which will be justified in the section below.

Through the overview function *File Profiler,* the words in the coursebook corpus are assigned different colours depending on which word frequency band they belong to, facilitating the visual analysis of the texts. Furthermore, each base word category shows its percentage score of appearance in the complete corpus, from which the word types and tokens can be seen. For a detailed view regarding coverage, the function *Statistics* shows the full specification of the base word lists and their type-token ratio, as well as the percentage of complete token coverage, i.e., lexical coverage (Anthony, 2022).

As mentioned in the theoretical section, the *English Vocabulary Profile* is not a corpus, but rather an online resource which has divided 15696 and 15389 words of British and American English, respectively, into the six threshold levels formulated by the CEFR (Council of Europe, 2020, p. 36). The division of these words is based on the annually updated, multi-billion-word Cambridge Learner Corpus, representing EFL learner-produced language as seen through examination scripts, exercises from coursebooks, and other materials used in EFL classrooms around the world (Cambridge University Press, 2015a). Despite the fact that the corpus is said to include both British and American English, as well as other varieties, the division in the *English Vocabulary Profile* only takes British and American English into consideration (Cambridge University Press, 2015c). Nevertheless, this division matches the representation of the BNC/COCA list in this regard, thus, facilitating the comparison between word frequencies and threshold levels. Thus, from the *File Profiler*, a selection of words was chosen and looked up on the website *English Vocabulary Profile Online* (Cambridge University Press, 2015c). The aim of this is to explain the origin of Nation's suggestion that the goal knowledge of a CEFR level B1-B2 student corresponds to 4,000 word families including technical vocabulary (Victoria University of Wellington, n.d.A).

4.2.2 Corpus Lists

As seen in the literature review, a majority of the corpus studies previously carried out have chosen either to use RANGE (Victoria University of Wellington, n.d.B) with GSL (Smith, 2023) or BNC/COCA lists (Anthony, n.d.). During the process of selecting word lists for this study, several limitations have been identified. Overall, the choice would have been better aided by lists representing the global varieties of English, since many of these are often depicted in the Swedish upper secondary coursebooks. Notwithstanding, the choice fell upon BNC/COCA, mainly because these corpora represent both American English and British English, but also since these are the varieties used for reference in the *English Vocabulary Profile* (Cambridge University Press, 2015c).

Another plausible impediment lies in the fact that both the BNC and COCA corpora represent NS-produced language (COCA, n.d.; BNC, n.d.), whereas *English Vocabulary Profile* (Cambridge University Press, 2015c) has made its division based on EFL student-produced language (Cambridge University Press, 2015a), by which a possible discrepancy in correspondence is highly likely to be found. Accordingly, GSL would have been a natural choice since it represents a language of practical use for EFL students. Yet despite its new version, which also gives it the highest score on recency, the New General Service List (Charlie

Browne Company Inc., 2023), only displays American English, which does not correspond to the division made in *English Vocabulary Profile* (Cambridge University Press, 2015c) and can therefore not give the necessary width needed for the comparison of research question three. Thus, even though the most recent update of COCA is 2019 (COCA, n.d.), and of BNC 2007 (University of Oxford, 2022), the BNC/COCA list has the closest approximation to the criteria for selection.

4.2.3 Coursebook Corpus

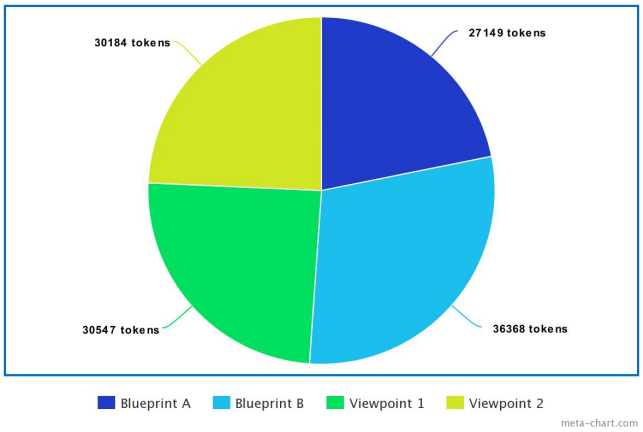
After gaining access to online copies of the four textbooks from the publishers, the target texts were converted from PDF to Word files. Prior to carrying out the analysis to answer the research questions, the accuracy of text conversion needed to be tested. The rationale behind this was that if a word has been incorrectly converted it would most likely be registered as '*not in lists*', and consequently, show incorrect results of the word frequencies. Consequently, all four documents were read through to spot spelling errors, aided by a test-run through *AntWordProfiler* (Anthony, 2022b), during which process also randomly, and incorrectly hyphenated words were found, such as for example *bodygu- ard, raci- al, physi- cal, enhance- ing*. During this stage, it was also discovered that the software does not differentiate between capital and minuscule letters, registering them as france, katrina, etcetera. Moreover, to facilitate the use of *AntWordProfiler* (Anthony, 2022), all text formats subsequently needed to be changed into UTF8, ending all document names in *.txt.*

Once all formatting was completed, a corpus was created consisting of four documents ranging between approximately 27,000-30,000 running words, also called tokens, totalling 124,248 tokens. These four documents represented the coursebooks *Blueprint A-B* (Lundfall & Nyström, 2017; 2018) and *Viewpoints 1-2* (Gustafsson & Wivast, 2017; 2018). Of these four coursebooks the text excerpts in *Blueprint A* contained the smallest number of tokens, whereas *Blueprint B* contained the largest. The content ratio between *Viewpoints 1-2* was almost equal (see Figure 2-3).

Figure 2. Token count in books and corpus in total.

	Token Count
Blueprint A	27,149
Blueprint B	36,368
Viewpoints 1	30,547
Viewpoints 2	30,184
	Total: 124,248

Figure 3. Token count in books.



The four coursebooks vary slightly in size, which might affect the presentation of both the contents of the word families, especially the first 3,000, as well as the repetition of words between the books. However, if a book contains a smaller number of tokens, it may still contain a higher number of academic texts, such as articles, for example. Therefore, accounting for the content of the corpus in terms of the distribution of both the genres and tokens is decisive for the analysis of the results.

4.3 Method and Analysis

The aforementioned selection of coursebooks, corpus lists, software and website tool should accumulate the necessary data to answer the research questions. The first research question (RQ1) aims to accumulate data regarding the representation of the first three word frequency bands and the extent to which words are recycled within these. The second research question (RQ2) is aimed at analysing the possible progression of word frequencies and the recycling of words beyond the first three word frequency bands. To answer the first two questions, *AntWordProfiler* (Anthony, 2022b) will be used. The third research question (RQ3) regarding

the comparison between word frequency bands and the threshold division of words will be answered through *English Vocabulary Profile* (Cambridge University Press, 2015c).

5 Results

5.1 Representation of the First Three Word Frequency Bands

Regarding the representation of the first 3,000 word families, the intersection of *basewrd 3* and *token_cum%* in the table shows the total percentage, which can also be interpreted as the required lexical coverage for this level. The total percentage for each word frequency band respectively, are shown in the three boxes under *token_count%*. The number of words to which these correspond is shown under *token_count,* and the number of words without repetition included is shown under *type_count.*

Figure 4, Blueprint A.

	file_name	list_id	token_count	token_count_%	token_cum_%	type_count
5	basewrd1	5	22008	79,5	79,51	1900
6	basewrd2	6	2137	7,72	87,23	1056
7	basewrd3	7	1112	4,02	91,25	673

Figure 5, Blueprint B.

	file_name	list_id	token_count	token_count_%	token_cum_%	type_count
4	basewrd1	4	29863	80,51	80,51	2156
5	basewrd2	5	2570	6,93	87,44	1319
6	basewrd3	6	1212	3,27	90,71	764

Figure 6, Viewpoints 1.

	file_name	list_id	token_count	token_count_%	token_cum_%	type_count
4	basewrd1	4	26400	83,74	83,74	1869
5	basewrd2	5	1908	6,05	89,79	914
6	basewrd3	6	786	2,49	92,29	463

Figure 7, Viewpoints 2.

	file_name	list_id	token_count	token_count_%	token_cum_%	type_count
4	basewrd1	4	25034	81,47	81,47	2053
5	basewrd2	5	2048	6,67	88,14	1141
6	basewrd3	6	950	3,09	91,23	654

As shown in figure 4-7, the total representation of the first 3,000 word families in all four books ranges between 90-92%. For the representation of words in the first frequency band the percentage shows between 79-84%, for the second between 6-8%, and for the third 2-4%, respectively, indicating a rapid decrease in all frequency bands throughout all four books. A similar diminishing speed is shown in the last column, *type_count,* which, when compared to the *token_count,* hints at the frequency of repetition.

5.2 Word Repetition

Departing from the need for a word to be repeated at least 10 times in order for the acquisition to take place (Pellicer-Sanchez & Schmitt, 2010, p. 42; Webb, 2007, p. 60), words repeated at a lower rate than this represent insufficient exposure for word retention. However, in the first frequency band only the first 336 words of *Blueprint A,* 338 from *Blueprint B,* 370 from *Viewpoints 1,* and 346 words from *Viewpoints 2* were repeated at least 10 times. Simultaneously, many of the words in the first frequency band are function words, such as *the, and, a, of, to,* etc. (Nation, 2013, p. 18), and the rest of the words in this frequency band appear at such a high frequency in general language use, so at the learning stage reached in the upper secondary, the acquisition of these words is likely to have been taken care of much earlier. Consequently, with regards to the repetition of words, a closer look will instead be taken at the type-token ratio and the effects of the rapid decrease of word representation found in the second and third frequency bands.

As previously mentioned, tokens are the total amount of words in a text, whereas types are the number of different words in it. Regarding the repetition of words 10 times or more in the second and third frequency band, calculations landed on the following.

Figure 8. The second frequency band.

	Tokens	**Types**	**Words repeated 10 times or more**	**Calculation**	**Percentage**
Blueprint A	2137	1056	15	15/1056 =	**0,9%**
Blueprint B	2570	1319	15	15/1319 =	**1%**
Viewpoints 1	1908	914	11	11/914 =	**1%**
Viewpoints 2	2048	1141	6	6/1141 =	**0,5%**

Figure 9. The third frequency band.

	Tokens	Types	Words repeated 10 times or more	Calculation	Percentage
Blueprint A	1112	673	7	7/673 =	**1%**
Blueprint B	1212	764	6	6/764 =	**0,7%**
Viewpoints 1	786	463	3	3/463 =	**0,6%**
Viewpoints 2	950	654	1	1/654 =	**0,1%**

The calculations show that the percentage of words repeated 10 times or more in either frequency band does not exceed 1% in any of the four coursebooks, indicating a low recycling of words beyond the first frequency band. As was suspected from the proximity between tokens and types, there was not much room for repetition of types. Despite the very low frequencies of repetition throughout all four books, the majority of the vocabulary in the first three frequency bands may already have been acquired, owed to their general frequency of appearance. If not, these books are coursebooks and come equipped with vocabulary lists and exercises, if need be. However, the extent of use of EFL coursebooks in the Swedish context differs from classroom to classroom.

5.3 Progression of Word Frequency Bands and Repetitions

Before zooming in on how the word frequency bands progress between books, an overall view will be given in terms of their total representation in all four books (cf. Appendices 1-4). Uniformly, all books contain base words from all 34 available base word lists. Yet, as expected for all books, the TTR (Type Token Ratio) at this level is very low, but with a sudden increase in the 31st frequency band, as well as a significant number of off-list words. In each table below, the numbers represent type/tokens in the 31st and 34th frequency band, as well as words excluded from categorisation according to the BNC/COCA list.

Figure 10. Type token ratio.

Frequency band	31st	34th	Not in lists
Blueprint A	269/860	15/19	116/211
Blueprint B	292/909	21/25	178/373
Viewpoints 1	189/625	7/8	95/227
Viewpoints 2	192/642	12/14	152/274

Since between 95-178 words in all books appear as *not in lists,* this hints that either these words are even farther above in terms of frequency, or, they have not been categorised. Some examples of the words in this category are *microplastic, biohackers, crispr, blacklivesmatter, hashtag, instagram, blazin, drippin, Ayesha, kretowicz, ewa.* Looking closer at these words, a possible categorisation is specialised vocabulary, representing online media, colloquial variations, and names whose spelling has been considered 'unusual'. A similar categorisation seems to be the case for the words of the 33rd frequency band, containing words such as *runaway, online, girlfriend, checkout, download, laptop.* The appearance of these types of words in such a low-frequency band questions the categorisation of words in the COCA and BNC (n.d.; n.d.), since one would think that these would be of more frequent appearance in today's general use of language, or at least categorised as more relevant to modern language.

The small number of words in the 34th frequency band seem to represent acronyms, with examples such as *cv, fda, org, uk, pms, cpu, phd, ceo,* whereas the unexpected increase of the words in the 31st frequency band reflects proper nouns, with examples such as *Kate, London, Toby, Derek,* etc. From this word frequency band, between 2-3% per cent and 4-5% correspond to tokens and word types, respectively, of each coursebook (c.f. Appendices 1-4). A possible explanation for this is found in what Nation (2013, p. 27) described in the categorisation of the BNC (n.d.), i.e., that a substantial number of low-frequency words are proper nouns, but no account is made of what lies behind the categorisation of words appearing in the *not in lists.*

Turning to the progression of word frequency bands, as a gauge for comparison, previous studies have searched for at which word frequency bands a lexical coverage level of 95-98% is reached (e.g., Sun & Dang, 2020; Alcaraz, 2009). As stated in the introduction of the answer to the first research question, the percentage of the total representation of word frequencies up to a certain level is shown in the column called *token_cum,* also called lexical coverage. With the purpose of avoiding long tables, seeing that all four books simultaneously reach a 98% coverage level at the 31st frequency band, a smaller table will be used to summarise the coverage levels of 95% for all four books (cf. figure 11). For an overview of both 95% and 98% lexical coverage scores, see Appendices 1-4.

Figure 11. 95% coverage level for all four books.

Word frequency band	***Blueprint A***	***Blueprint B***	***Viewpoint 1***	***Viewpoint 2***
1000	79.5	80.51	83.74	81.47
2,000	87.22	87.44	89.79	88.14
3,000	91.24	90.71	92.29	91.23
4,000	92.48	92.42	93.59	92.83
5,000	93.45	93.48	94.44	93.87
6,000	93.98	94.18	**95.02**	94.48
7,000	94.4	94.66	95.3	94.92
8,000	94.66	94.99	95.62	**95.32**
9,000	94.86	**95.21**	95.93	95.57
10,000	94.97	95.36	96.11	95.75
11,000	**95.05**	95.49	96.22	95.9
12,000	95.09	95.62	96.26	95.97

The table shows that *Viewpoints 1* is first at reaching a 95% coverage at the 6th word frequency band, whereas *Blueprint A* scores the highest at the 11th word frequency band. *Blueprint B* and *Viewpoints 2* are in the middle at the 9th and 8th word frequency bands, respectively. From this can be seen that a natural progression in word frequency bands exists between *Viewpoint 1-2,* yet interestingly, the opposite is the case for *Blueprint A-B,* indicating a higher presence of mid- and low-frequency words in book A. Looking back at the division of genres in figure 1 on page 17, even though book B contains a higher number of texts in total, book A contains more articles than book B, which might be a potential explanation for a higher presence of specialised vocabulary in book A. Moreover, this is also surprising since *Blueprint A,* according to figures 2-3 on page 21, contains the smallest number of tokens in total.

Regarding the repetition from the fourth frequency band and up, after the dramatic decline seen in the second and third frequency bands, all occurrences in all books remain low. Thus, the degree of retention cannot be dependent on extensive reading only, the vocabulary lists and exercises will also need to be used in order to ensure vocabulary acquisition. In each book, barely one word is repeated more than 10 times after the 7th frequency band. Although, interestingly, in the 31st-word frequency band representing proper nouns, all four books showed repetitions of 10 times or more in between 9-18 words.

5.4 Frequency Bands versus Threshold Levels

Turning to the data regarding the answering of the last research question investigating the potential correspondence between frequency bands and the threshold levels, words will now be selected from their appearance in the word frequency lists found in all four coursebooks. A previous attempt made by Nation suggested that the B1 and B2 level used in the CEFR (Council of Europe, 2020, p. 36) corresponds to a knowledge of the first 2,000-3,000 and 4,000 word families, respectively, plus "1000-2000 words of specialised vocabulary" at the B2-level (Victoria University of Wellington, n.d.A). The origins are not accounted for in the suggestion, however, based on the numbers presented, we can assume that they correspond to a text coverage of 95%, as suggested in Nation (2013, p. 26) and Nation (2006, p. 70-72). However, since the second frequency band does not correspond to a sufficient amount of the most frequent vocabulary, the words for selection will be taken from the third and fourth frequency bands only. The frequency of types from these bands is as follows.

Figure 12. Word types in the 3rd and 4th frequency bands.

	3rd Frequency Band	**4th Frequency Band**
Blueprint A	673	262
Blueprint B	764	455
Viewpoint 1	463	262
Viewpoint 2	654	368

To avoid the workload of looking up all types, every 100th word will be chosen. A total representation of 4 words from each book will be chosen, thus, obtaining extra options from the books whose frequency bands contain up to 600-700 types. This will also remedy possible appearances of the same word from different books. Since both *Viewpoint 1* and *Blueprint A* only have 262 types in the fourth frequency band, their last appearing word in the fourth frequency band will also be selected.

Figure 13. Words selected from the 3rd and 4th frequency bands.

	3rd Frequency band	**4th Frequency band**
Blueprint A	*media, solutions, shrugged, relatives*	*devil, escorting, untucked, shallow*
Blueprint B	*violence, apologise, authors, occupation*	*pencil, trousers, poster, kneeling*
Viewpoint 1	*text, privileged, rebellion, sculpt*	*hood, rugs, lively, spectacle*
Viewpoint 2	*gender, alert, hosting, philosopher*	*thy, plunge, optimistic, amateur*

After extracting the words above from all four books, these were subsequently searched for on the *English Vocabulary Profile* webpage (Cambridge University Press, 2015c). As expected, the results on the webpage showed only the unconjugated base form of each search word, hence, the same format will be used in the results shown in the table below. Furthermore, since the results represented a significantly wider categorisation than just between the expected B1-B2, the whole threshold scale will be included in the table, also adding an extra box for words not found in the *English Vocabulary Profile* (Cambridge University Press, 2015c). For an easier overview of the level division of words, and similar to the reference made between the six rainbow colours and the threshold levels in the CEFR (Council of Europe, 2020, p. 36) the same rainbow colours will be used, and words not found will be marked without colour (cf. figure 14-15).

Figure 14. Word division into frequency bands and threshold levels.

A1	**A2**	**B1**	**B2**	**C1**	**C2**	**Not found**
	3rd Frequency Band			**4th Frequency Band**		
Blueprint A	***media, solutions, shrugged, relatives***			***devil, escorting, untucked, shallow***		
Blueprint B	***violence, apologise, authors, occupation***			***pencil, trousers, poster, kneeling***		
Viewpoints 1	***text, privileged, rebellion, sculpt***			***hood, rugs, lively, spectacle***		
Viewpoints 2	***gender, alert, hosting, philosopher***			***thy, plunge, optimistic, amateur***		

Before investigating the possible parallels between frequency bands and threshold levels, a closer look will be taken at the general representation of threshold levels in the coursebooks. As stated earlier, the equivalence of the Swedish upper secondary EFL courses 5-6 is met at the B1-B2 levels in the CEFR (Skolverket, 2022, p. 7). Accordingly, the coursebooks *Blueprint A* and *Viewpoints 1,* designated for the English 5 courses, would be expected to contain mainly words up to B1 level, whereas the coursebooks *Blueprint B* and *Viewpoints 2,* designated for English 6, would contain mainly words up to B2-level, and none of the books would contain words from the two C-levels. Be that as it may, interestingly, the division seems to span over the entire threshold scale, with *Blueprint B* giving the widest representation, and *Viewpoints 2* the narrowest.

To answer research question three concerning the comparison between threshold levels and frequency bands, the wide distribution of threshold level division between frequency bands 2-3 is striking. Words from both B1 and B2 are represented in both frequency bands, words from A1 are present in the fourth, and words from C2 are present in both frequency bands. Such

a scattered division confirms the suspicion of the incompatibility between the B1-B2 threshold level and knowledge of the first 2,000-4,000 word families plus specialised vocabulary.

Naturally, there are several limitations questioning these results. First, the selected words are only 32 types from a corpus with 124,248 tokens, thus, only describing the division of a very small number of words. Second, several words represent different threshold divisions depending on the word meaning referred to; a table showing all the different meanings represented within these words can be seen in Appendix 5. Last, that coursebooks should only contain one type of word, for example, only B1 words, is not only impossible, but also, does not provide sufficient pedagogical challenges for vocabulary acquisition to take place. Simply put, the wider the exposure, the wider the gains.

6 Discussion

To summarise the results, we will look at the data found regarding RQ1-3 in chronological order. Simultaneously, each research question will be related to similar issues previously found in research. Thereafter, pedagogical implications will be accounted for and the paper closes with suggestions for future research.

6.1 Coverage of the First 3,000 Word Families and Recycling

Regarding RQ1, the overall representation of the first three frequency bands in all four books ranges between 90-92%, of which, in the first frequency band the majority of tokens were covered between 79-84%. Of the representation found in the second and third frequency band, a sharp decline was identified, which correlated with decimated word repetition, indicating low recycling of words beyond the first frequency band (cf. Matsuoka & Hirsh, 2010, p. 65; Nordlund & Norberg, 2020, p. 98; Bergström, Norberg & Nordlund, 2022, p. 12). Notwithstanding, on the one hand, in accordance with the conclusion of Norberg and Nordlund, a text comprising a levelled vocabulary in terms of word frequency presentation is unnatural (2018, p. 469). On the other hand, if the frequency of repetition would have been investigated in each individual text excerpt, chances are high that repetition of specialised vocabulary would be dense enough to promote vocabulary acquisition since the majority of the texts in the coursebooks deal with a particular theme.

Furthermore, since many of these themes are chosen because of their contribution to the covering of themes in the syllabi, such as sustainable development, scientific texts, and cultures of different English-speaking areas, to mention a few (Skolverket, 2020), chances are high that

the words will be repeated in additional classroom material. Yet, as stated earlier, the use of a coursebook in the Swedish classroom varies, thus, the extent to which both extra material and coursebooks are used cannot be determined, whereby explicit vocabulary teaching of relevant terminology is recommended (Matsuoka & Hirsh, 2010, p. 65).

In conclusion, if the texts in EFL coursebooks would be composed to reflect a levelled vocabulary in terms of sufficient representation of the first 3,000 word families, as well as repeating the majority of all words 10 times or more, they would most certainly look odd. Thus, in order to fulfil the required exposure of a minimum of 3,000 word families (Schmitt & Schmitt, 2012, p. 498) and sufficient repetitions, additional classroom material has to be used, such as the reading of novels.

6.2 Total Representation of Frequency Bands and Progression

In relation to RQ2, disparate distribution of a 95% lexical coverage between the 6th and 11th frequency band was found when comparing all books, and a linear progression of vocabulary load was only identified between *Viewpoints 1-2,* while the opposite was the case for *Blueprint A-B.* A possible reason for this may be that even though *Blueprint B* (Lundfall & Nyström, 2018) contains a substantially higher number of texts, *Blueprint A* (Lundfall Nyström, 2017) contains a higher number of news articles, which, thus, may contain a higher density of specialised vocabulary (Nation, 2013, p. 30). As with the results for RQ1, the presence of specialized vocabulary may also be an effect of texts whose themes are dealing with certain topics, requiring more advanced language, as outlined by Skolverket (2022). Be that as it may, since the content and progression of advanced vocabulary were not linear between the coursebooks of this study, just as seen for Sun and Dang (2020, p. 7), the 95% coverage level can be questioned. Adding to this, a lower coverage should be accepted due to the fact that coursebooks come accompanied with vocabulary lists and exercises, and also because vocabulary acquisition is expected to take place during courses. After all, the levels 95 and 98% were not based on coursebook studies, but rather on texts in general, implying that a 95% coverage can merely be seen as a goal, the reaching of which only can be confirmed or refuted for individual students during reading comprehension tests, such as those carried out by Alsaif and Milton (2012, p. 32). Yet, the importance of reaching a 95% comprehension level when reading is not confirmed by either Skolverket (2022) or the CEFR (Council of Europe, 2020) since the more dynamic term ‘skill’ is preferred instead of fixed ‘knowledge’ (van Ek, 1976, p. 5).

All books simultaneously reached a 98% lexical coverage at the 31st frequency band due to a sudden and almost identical increase of word types represented by proper names, in accordance with the categorisation of proper names in the BNC Corpus (Nation, 2013, pp. 21-22, p. 27). Moreover, no significant number of words of the less frequent bands were repeated at a significant rate, just as found in previous studies (c.f. Matsuoka & Hirsh, 2010, p. 65; Nordlund & Norberg, 2020, p. 98; Bergström, Norberg & Nordlund, 2022, p. 12), whereby they will be excluded from further analysis.

Beyond the first 31 frequency bands, a substantial number of words and names were excluded from proper categorisation or were categorised as *not in lists,* thus, questioning the system for categorisation made in the BNC/COCA list (Anthony, n.d.). Alternative categorisation for these words might be specialised vocabulary, online media, colloquial variations, and names of which spelling has been considered 'unusual' (cf. p. 23). Similarly, when using the GSL, O'Loughlin concluded (2012, p. 263) that the absence of modern words questions the validity of the categorisation made in the corpus list. In this study, the words excluded from, or beyond proper categorisation, seen in section 5.3, pages 24-25, were categorised as belonging to frequency bands 31 and up. Based on previous suggestions where a 95% coverage reflects a knowledge of 5,000 word families (Laufer 1989, p. 317) and a 98% coverage reflects a knowledge of 8,000-9,000 word families (Nation, 2006, pp. 71-72), such a categorisation suggests that words such as *Instagram, girlfriend, drippin,* or *laptop* are way beyond what students need to know, or already do know, which also questions the validity of the categorisation made in corpus lists.

Apropos modern language use, contrasting to when the extensive vocabulary list of Thorén was compiled (1976), the difficulty of mapping the language use of the students today, for example online, results in unidentified needs of vocabulary development. Although it may be expected that many of the words in frequency bands 33 and up are already known to students, section 5.3 pages 24-25, just as those in the first three frequency bands, their placement at such a low frequency further discredits the usefulness of word frequencies to be responsible for measuring the readability of a text.

Similar to the conclusion reached about RQ1, to compensate for insufficient representation of word frequencies and the repetition of their words, additional classroom materials are needed, as well as explicit vocabulary teaching. However, the unelucidated use and need of vocabulary for the EFL students of today challenges previously formulated goals and recommendations for development.

6.3 Lexical Coverage versus Threshold Levels

In the data found regarding RQ3, a possible incongruity in attempting to correlate the threshold levels to word frequency bands was identified, since the words from the threshold levels were scattered haphazardly between frequency bands 3-4 (cf. figure 14, p. 28). Not only were some words possible to place into several threshold levels depending on which semantic meaning of the word is referred to, but also the haphazard placement of words in different levels when compared to the categorisation of word frequencies suggests a poor correspondence of word frequencies and threshold levels. However, the content of categorised words in the *English Vocabulary Profile* (Cambridge University Press, 2015c) barely reaches 16,000 words, whereas the corpus list BNC/COCA (Anthony, n.d.) is based on multi-million-word corpora. Such imbalanced word content is sure to complicate comparisons.

Furthermore, based on the questioning of the validity of the categorisation of words in frequency bands seen in section 6.2, page 31, looking at words one might consider of modern use as pertaining to frequency bands 31 and up indicates that words in this categorisation are likely to be part of students' vocabulary, although such a high categorisation indicates the opposite. Similar results were found by O'Loughlin (2012), who used the GSL list where words such as *plough* and *shilling* (p. 263) were categorised as frequently used words. In conclusion, the inadequacy of using word frequencies as a recommendation for the trajectory of vocabulary acquisition for EFL students is thus confirmed.

7 Pedagogical Implications and Suggestions for Future Research

Even though a vast representation of frequency bands was found in all four coursebooks, the low recycling of words beyond the first frequency band indicates that vocabulary acquisition requires additional reading materials. Since the focus of coursebook designers probably will continue to prioritise content rather than vocabulary load, coursebooks will always be abaft the development of the syllabus, and the extent to which they are used will therefore keep fluctuating. In this study, it has been shown that the specialised content of a great number of texts reflect an advanced vocabulary in terms of word frequencies, which, together with additional teaching materials would probably promote adequate vocabulary gains. However, since the word frequency progression identified between books in this study was not linear, one can question their role as indicators for in which order vocabulary should be acquired.

Furthermore, due to the incompatibility of word frequencies and threshold levels, additional ways of gauging what is the appropriate vocabulary to learn need to be found through studies investigating a greater number of words from EFL coursebooks. Such studies need to be based on not only learner language instead of NS language but also on drawing parallels between Swedish and English language use. Therefore, prior to an initiative to update Thorén (1976), a collection of data regarding the average vocabulary uses and needs of EFL students of the upper secondary of Sweden is needed. Not until this is acquired, can we properly formulate goals and recommendations for vocabulary development, after which, updates can, and will need to be carried out at a much higher rate than previously seen, owing to the rapid digitalisation we have experienced over the past 50 years. Said studies, updates and reformulations can thus, assist EFL teachers to choose material appropriate for the vocabulary development of their students in the Swedish context.

References

Alcaraz, G. (2009). Frequency and Functionality: Two Keys for L2 Coursebooks. *International Journal of English Studies, 9*(3), 61-72. https://revistas.um.es/ijes/article/view/99521

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *The Language Learning Journal, 40*(1), pp.21–33. https://doi.org/10.1080/09571736.2012.658221

Anthony, L. (2022). AntWordProfiler (Version 2.0.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software/antwordprofiler/

Anthony, L. (n.d.). *BNC/COCA family lists + extras (Version 2.00).* In *Laurence Anthony.* Available from https://www.laurenceanthony.net/software/antwordprofiler/

Bergström, D., Norberg, C. & Nordlund, M. (2022). Do textbooks support incidental vocabulary learning? – a corpus-based study of Swedish intermediate EFL materials. *Education Inquiry*, (*no issue*), pp.1-19. https://doi.org/10.1080/20004508.2022.2163050

Bergström, D., Norberg, C. & Nordlund, M. (2023): "The Text Comes First" – Principles Guiding EFL Materials Developers' Vocabulary Content Decisions. *Scandinavian Journal of Educational Research, 67*(1), pp.154-168. https://doi.org/10.1080/00313831.2021.1990122

British National Corpus. (n.d.). Retrieved 2023-04-06 from https://www.english-corpora.org/bnc/

Brown, H. D. & Lee, H. (2015). *Teaching by Principles.* Pearson Education.

Browne, C. (2013). The New General Service List: Celebrating 60 years of vocabulary learning. *The Language Teacher, 37*(4), pp.13–16. https://jalt-publications.org/files/pdf-article/37.4tlt_featureds.pdf

Cambridge Dictionary. (2023). Corpus. In *Cambridge Dictionary.* Retrieved 27-04-2023 from https://dictionary.cambridge.org/dictionary/english/corpus

Cambridge University Press. (2015a) *Compiling the EVP.* Retrieved 2023-05-10 from https://www.englishprofile.org/wordlists/compiling-the-evp

Cambridge University Press. (2015b). *English Profile – what the CEFR means for English.* Retrieved 2023-04-06 from https://www.englishprofile.org

Cambridge University Press. (2015c). *English Vocabulary Profile Online*. Retrieved 2023-04-04 from https://www.englishprofile.org/wordlists/evp

Charlie Browne Company Inc. (2023). *New General Service List Project.* https://www.newgeneralservicelist.com/

Cobb, T. (2017). *Families v. Lemmas.* Lextutor. Available from https://www.lextutor.ca/familizer/define.html#:~:text=it%20for%20free.-,',from%20different%20parts%20of%20speech.

Cobb, T. (2009). VocabProfile. (Version VP English v.3). [Computer Software]. Available at: http://www.lextutor.ca/vp

Cobb. T. (n.d.) *Compleat Lexical Tutor.* (Various versions). Retrieved 2023-04-06 from https://www.lextutor.ca/

Cook, V. (2016). *Second Language Learning and Language Teaching.* Routledge.

Corpus of Contemporary American English. (n.d.). Retrieved 2023-04-06 from https://www.english-corpora.org/coca/

Council of Europe (2020), *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, Strasbourg. Available at https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, (*34*)2, pp. 213-238. https://doi.org/10.2307/3587951

Ellis, N. (2002). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition, 24*(2), pp.143-188. https://doi.org/10.1017/S0272263102002024

Goulden, R., Nation, P., & Read, J. (1990). How Large Can a Receptive Vocabulary Be? *Applied Linguistics, 11*(4), pp.341-363. https://doi.org/10.1093/applin/11.4.341

Gustafsson, L., & Wivast, U. (2017). *Viewpoints 1.* (Upplaga 2). Gleerups.

Gustafsson, L., & Wivast, U. (2018). *Viewpoints 2.* (Upplaga 2.5). Gleerups.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8(*2), pp.689-696. https://doi.org/10125/67046

Hu, M., & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), pp.403–430. 13_1_10125_66973_rfl131hsuehchao.pdf

Hutchinson, T., & Torres, E. (1994). The textbook as agent of change. *ELT Journal*, *48*(4), pp. 315–328. https://doi.org/10.1093/elt/48.4.315

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Leech, G., Rayson, P. & Wilson, A. (2001). *Frequency Lists.* Available from https://ucrel.lancs.ac.uk/bncfreq/flists.html

Lundfall, C., & Nyström, R. (2017). *Blueprint A : Engelska 5* (Upplaga 3). Liber.

Lundfall, C., & Nyström, R. (2018). *Blueprint B : Engelska 6* (Upplaga 3). Liber.

Matsuoka & Hirsh. (2010) Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language, 22*(1), pp. 56–70. https://nflrc.hawaii.edu/rfl/item/208

McKay, S. (2006). *Researching Second Language Classrooms.* Routledge.

Nation, P. (2006). How Large a Vocabulary Is Needed for Reading and Listening? *The Canadian Modern Language Review, 63*(1), pp. 59-81 https://doi.org/10.1353/cml.2006.0049

Nation, P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language, 27*(1), pp.136–145. https://doi.org/10125/66705

Nation, I. (2013a). Testing vocabulary knowledge and use. In *Learning Vocabulary in Another Language* (Cambridge Applied Linguistics, pp. 514-568). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656.015

Nation, P. (2013b). The Goals of Vocabulary Learning. In *Learning Vocabulary in Another Language* (Cambridge Applied Linguistics, pp. 9-43). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656.003

Norberg, C., & Nordlund, M. (2018). A corpus-based study of lexis in L2 English textbooks. *Journal of Language Teaching and Research, 9*(3), 463–473. https://doi.org/10.17507/jltr.0903.03

Nordlund, M., & Norberg, C. (2020). Vocabulary in EFL teaching materials for young learners. *International Journal of Language Studies, 14*(1), pp.89–116. https://www.diva-portal.org/smash/get/diva2:1385144/FULLTEXT01.pdf

O'Keeffe, L. (2013). A framework for textbook analysis. *International Review of Contemporary Learning Research, 2*(1), pp.1-13. http://dx.doi.org/10.12785/IRCLR/020101

O'Loughlin, R. (2012). Tuning In to Vocabulary Frequency in Coursebooks. *RELC Journal 43*(2) pp.255 –269. https://doi.org/10.1177/0033688212450640

Oxford English Dictionary. (2023). *About.* Available at https://public.oed.com/about/#

Pellicer-Sanchez, A., & Schmitt, N. (2010). Incidental Vocabulary Acquisition from an Authentic Novel: Do Things Fall Apart? *Reading in a Foreign Language, 22*(1), pp.31-55. https://nflrc.hawaii.edu/rfl/item/207

Sheldon, L. (1988). Evaluating ELT textbooks and materials. *ELT Journal, 42*(4), pp.237-246. https://doi.org/10.1093/elt/42.4.237

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), pp.329-363. https://doi.org/10.1177/1362168808089921

Schmitt, N. and Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), pp.484 – 503. https://doi.org/10.1017/S0261444812000018

Skolverket. (2020). *[English - Syllabus].* https://www.skolverket.se/download/18.7f8c152b177d982455e1158/1615808938264/%C3%84mnesplan_engelska.pdf

Skolverket. (2022). *Kommentarmaterial till ämnesplanerna i moderna språk och engelska – gymnasieskolan och vux.* https://www.skolverket.se/publikationer?id=9918

Smith, S. (2023). *The General Service List (GSL).* Available from https://www.eapfoundation.com/vocab/general/gsl/

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System, 93*, pp.1-13. https://doi.org/10.1016/j.system.2020.102279

Thorén, B. (1976). *10000 ord för tio års engelska : Ordlista.* (2nd ed.). LiberLäromedel.

University of Oxford. (2022). *What is the BNC?* http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro

van Ek, J. (1976). *The Threshold Level – for Modern Language Learning in Schools.* Longman Group.

Victoria University of Wellington. (n.d.A). *The CEFR Levels, Word Parts and Vocabulary Sizes.* https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

Victoria University of Wellington. (n.d.B). *Vocabulary Analysis Programs.* https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

Webb, S. (2007). The Effects of Repetition on Vocabulary Knowledge. *Applied Linguistics, 28*(1), pp.46-65. https://doi.org/10.1093/applin/aml048

Word Frequency. (n.d.). *Word Frequency Data.* https://www.wordfrequency.info/coca.asp

Yang, L., & Coxhead, A. (2020). A corpus-based study of vocabulary in the New Concept English textbook series. *RELC Journal, 53*(3), 597–611. https://doi.org/10.1177/0033688220964162

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a Functionally Important Lexicon. *Journal of Reading Behavior*, *27*(2), pp.201–212. https://doi.org/10.1080/10862969509547878

Appendices

Appendix 1

Blueprint A

	file_name	list_id	token_count	token_count_%	token_cum_%	type_count	type_count_%	type_cum_%
4	basewrd1	4	22008	79,5	79,5	1900	38,1	38,1
5	basewrd2	5	2137	7,72	87,22	1056	21,18	59,27
6	basewrd3	6	1112	4,02	91,24	673	13,5	72,77
7	basewrd4	7	345	1,25	92,48	262	5,25	78,02
8	basewrd5	8	268	0,97	93,45	189	3,79	81,81
9	basewrd6	9	146	0,53	93,98	108	2,17	83,98
10	basewrd7	10	117	0,42	94,4	80	1,6	85,58
11	basewrd8	11	71	0,26	94,66	57	1,14	86,73
12	basewrd9	12	55	0,2	94,86	37	0,74	87,47
13	basewrd10	13	31	0,11	94,97	27	0,54	88,01
14	basewrd11	14	22	0,08	95,05	19	0,38	88,39
15	basewrd12	15	13	0,05	95,09	13	0,26	88,65
16	basewrd13	16	14	0,05	95,15	11	0,22	88,87
17	basewrd14	17	10	0,04	95,18	10	0,2	89,07
18	basewrd15	18	10	0,04	95,22	10	0,2	89,27
19	basewrd16	19	9	0,03	95,25	7	0,14	89,41
20	basewrd17	20	5	0,02	95,27	3	0,06	89,47
21	basewrd18	21	3	0,01	95,28	3	0,06	89,53
22	basewrd19	22	1	0	95,28	1	0,02	89,55
23	basewrd20	23	11	0,04	95,32	4	0,08	89,63
24	basewrd21	24	13	0,05	95,37	4	0,08	89,71
25	basewrd22	25	1	0	95,37	1	0,02	89,73
26	basewrd23	26	2	0,01	95,38	2	0,04	89,77
27	basewrd24	27	1	0	95,38	1	0,02	89,79
28	basewrd25	28	1	0	95,39	1	0,02	89,81
29	basewrd31	29	860	3,11	98,49	269	5,39	95,21
30	basewrd32	30	65	0,23	98,73	27	0,54	95,75
31	basewrd33	31	122	0,44	99,17	81	1,62	97,37
32	basewrd34	32	19	0,07	99,24	15	0,3	97,67
33	ignored	33	0	0	99,24	0	0	97,67
34	not_in_lists	34	211	0,76	100	116	2,33	100
35	TOTAL		27683	100		4987	100	

Appendix 2

Blueprint B

4	basewrd1	4	29863	80,51	80,51	2156	34,21	34,21
5	basewrd2	5	2570	6,93	87,44	1319	20,93	55,13
6	basewrd3	6	1212	3,27	90,71	764	12,12	67,25
7	basewrd4	7	632	1,7	92,42	455	7,22	74,47
8	basewrd5	8	396	1,07	93,48	267	4,24	78,71
9	basewrd6	9	258	0,7	94,18	202	3,2	81,91
10	basewrd7	10	180	0,49	94,66	146	2,32	84,23
11	basewrd8	11	121	0,33	94,99	88	1,4	85,63
12	basewrd9	12	83	0,22	95,21	71	1,13	86,75
13	basewrd10	13	53	0,14	95,36	51	0,81	87,56
14	basewrd11	14	50	0,13	95,49	38	0,6	88,16
15	basewrd12	15	46	0,12	95,62	38	0,6	88,77
16	basewrd13	16	28	0,08	95,69	27	0,43	89,2
17	basewrd14	17	17	0,05	95,74	13	0,21	89,4
18	basewrd15	18	14	0,04	95,78	13	0,21	89,61
19	basewrd16	19	6	0,02	95,79	6	0,1	89,7
20	basewrd17	20	8	0,02	95,81	7	0,11	89,81
21	basewrd18	21	12	0,03	95,85	3	0,05	89,86
22	basewrd19	22	3	0,01	95,85	3	0,05	89,91
23	basewrd20	23	3	0,01	95,86	3	0,05	89,96
24	basewrd21	24	2	0,01	95,87	2	0,03	89,99
25	basewrd22	25	2	0,01	95,87	2	0,03	90,02
26	basewrd23	26	1	0	95,87	1	0,02	90,04
27	basewrd24	27	0	0	95,87	0	0	90,04
28	basewrd25	28	0	0	95,87	0	0	90,04
29	basewrd31	29	909	2,45	98,33	292	4,63	94,67
30	basewrd32	30	74	0,2	98,53	23	0,36	95,03
31	basewrd33	31	149	0,4	98,93	114	1,81	96,84
32	basewrd34	32	25	0,07	98,99	21	0,33	97,18
33	ignored	33	0	0	98,99	0	0	97,18
34	not_in_lists	34	373	1,01	100	178	2,82	100
35	TOTAL		37090	100		6303	100	

Appendix 3

Viewpoints 1

	file_name	list_id	token_count	token_count_%	token_cum_%	type_count	type_count_%	type_cum_%
4	basewrd1	4	26400	83,74	83,74	1869	41,71	41,71
5	basewrd2	5	1908	6,05	89,79	914	20,4	62,11
6	basewrd3	6	786	2,49	92,29	463	10,33	72,44
7	basewrd4	7	411	1,3	93,59	262	5,85	78,29
8	basewrd5	8	269	0,85	94,44	189	4,22	82,5
9	basewrd6	9	183	0,58	95,02	103	2,3	84,8
10	basewrd7	10	86	0,27	95,3	67	1,5	86,3
11	basewrd8	11	103	0,33	95,62	67	1,5	87,79
12	basewrd9	12	98	0,31	95,93	49	1,09	88,89
13	basewrd10	13	57	0,18	96,11	36	0,8	89,69
14	basewrd11	14	34	0,11	96,22	25	0,56	90,25
15	basewrd12	15	13	0,04	96,26	9	0,2	90,45
16	basewrd13	16	95	0,3	96,56	17	0,38	90,83
17	basewrd14	17	12	0,04	96,6	8	0,18	91,01
18	basewrd15	18	4	0,01	96,62	4	0,09	91,1
19	basewrd16	19	5	0,02	96,63	4	0,09	91,19
20	basewrd17	20	9	0,03	96,66	8	0,18	91,36
21	basewrd18	21	4	0,01	96,67	4	0,09	91,45
22	basewrd19	22	2	0,01	96,68	1	0,02	91,48
23	basewrd20	23	10	0,03	96,71	3	0,07	91,54
24	basewrd21	24	3	0,01	96,72	2	0,04	91,59
25	basewrd22	25	0	0	96,72	0	0	91,59
26	basewrd23	26	0	0	96,72	0	0	91,59
27	basewrd24	27	0	0	96,72	0	0	91,59
28	basewrd25	28	3	0,01	96,73	1	0,02	91,61
29	basewrd31	29	625	1,98	98,71	189	4,22	95,83
30	basewrd32	30	75	0,24	98,95	28	0,62	96,45
31	basewrd33	31	96	0,3	99,25	57	1,27	97,72
32	basewrd34	32	8	0,03	99,28	7	0,16	97,88
33	ignored	33	0	0	99,28	0	0	97,88
34	not_in_lists	34	227	0,72	100	95	2,12	100
35	TOTAL		31526	100		4481	100	

Appendix 4

Viewpoints 2

	file_name	list_id	token_count	token_count_%	token_cum_%	type_count	type_count_%	type_cum_%
4	basewrd1	4	25034	81,47	81,47	2053	36,9	36,9
5	basewrd2	5	2048	6,67	88,14	1141	20,51	57,4
6	basewrd3	6	950	3,09	91,23	654	11,75	69,16
7	basewrd4	7	491	1,6	92,83	368	6,61	75,77
8	basewrd5	8	318	1,03	93,87	257	4,62	80,39
9	basewrd6	9	188	0,61	94,48	164	2,95	83,34
10	basewrd7	10	136	0,44	94,92	108	1,94	85,28
11	basewrd8	11	122	0,4	95,32	92	1,65	86,93
12	basewrd9	12	78	0,25	95,57	62	1,11	88,05
13	basewrd10	13	54	0,18	95,75	50	0,9	88,95
14	basewrd11	14	46	0,15	95,9	32	0,58	89,52
15	basewrd12	15	24	0,08	95,97	20	0,36	89,88
16	basewrd13	16	22	0,07	96,05	19	0,34	90,22
17	basewrd14	17	8	0,03	96,07	8	0,14	90,37
18	basewrd15	18	8	0,03	96,1	8	0,14	90,51
19	basewrd16	19	11	0,04	96,13	10	0,18	90,69
20	basewrd17	20	4	0,01	96,15	4	0,07	90,76
21	basewrd18	21	10	0,03	96,18	7	0,13	90,89
22	basewrd19	22	5	0,02	96,2	3	0,05	90,94
23	basewrd20	23	10	0,03	96,23	8	0,14	91,09
24	basewrd21	24	3	0,01	96,24	3	0,05	91,14
25	basewrd22	25	4	0,01	96,25	4	0,07	91,21
26	basewrd23	26	1	0	96,25	1	0,02	91,23
27	basewrd24	27	0	0	96,25	0	0	91,23
28	basewrd25	28	0	0	96,25	0	0	91,23
29	basewrd31	29	642	2,09	98,34	192	3,45	94,68
30	basewrd32	30	50	0,16	98,51	20	0,36	95,04
31	basewrd33	31	171	0,56	99,06	112	2,01	97,05
32	basewrd34	32	14	0,05	99,11	12	0,22	97,27
33	ignored	33	0	0	99,11	0	0	97,27
34	not_in_lists	34	274	0,89	100	152	2,73	100
35	TOTAL		30726	100		5564	100	

Appendix 5

Threshold Level Division of Words from the Coursebooks

	A1	A2	B1	B2	C1	C2	Not found
Blueprint A			*Solution, relative*	*The media, devil* (evil being), *shallow* (not deep)		*Shrug, devil* (badly-behaved person), *shallow* (not serious)	*Escort, untuck*
Blueprint B	*Pencil, trousers*	*Occupation* (job), *poster,*	*Apologise, author*	*Violence* (hurt), *kneeling*	*Occupation* (hobby)	*Violence* (extreme force), *Occupation* (control)	
Viewpoints 1		*Text* (verb; message),	*Text* (words; piece of writing), *rugs, lively*	*Text* (book/play), *Hood*	*Privileged* (advantage), *rebellion,*	*Privileged* (opportunity)	*Sculpt, spectacle*
Viewpoints 2				*Gender* (grammar; sex), *philosopher, optimistic*	*Alert* (noun; adjective), *hosting, plunge* (become lower), *amateur* (noun; no skill; hobby)	*Alert* (verb), *plunge* (phrase; phrasal verb; idiom)	*Thy,*