VILNIUS UNIVERSITY
FACULTY OF PHILOLOGY

Nurassyl Silan

English Studies (Linguistics)

A Corpus-Driven Study of Lexis in EFL Graded Readers in Kazakhstan

Master Thesis

Academic Supervisor: Assoc. Prof. Dr Rita Juknevičienė

Vilnius

2026

Anotacija

Šiame tyrime nagrinėjama septynių Kazachstane naudojamų anglų kalbos skaitymo knygelių, pritaikytų skirtingiems mokymosi lygiams, leksika. Tyrimo tikslas – įvertinti Oksfordo leidyklos parengtų skaitinių kalbos sudėtingumą ir tinkamumą vidurinės mokyklos mokiniams bei ištirti. ar šių leidinių žodynas atspindi bendrinę anglų kalbą. Tyrimas grindžiamas tekstynų lingvistikos metodais: jame analizuojamas autoriaus sudarytas skaitymo knygelių tekstynas. Tekstyną sudaro 246 969 žodžiai. Norint atsakyti į šį klausimą, tyrime derinami šie tyrimo metodai: atskirų žodžių analizė, kolokacijų analizė, žodyno profiliavimas ir skaitomumo analizė. Atskirų žodžių analizė sutelkta į penkiasdešimt dažniausių daiktavardžių, veiksmažodžių ir būdvardžių ir jų palyginimą su Britų nacionalinio tekstyno (BNC2014) grožinės literatūros patekstyniu. Atskirų žodžių analizė padeda įvertinti, kiek adaptuoti skaitiniai atspindi bendrąją anglų kalbos leksiką. Rezultatai rodo, kad veiksmažodžiai labiausiai sutampa su tekstyno duomenimis, o daiktavardžiai – mažiausiai. Priežastis, kodėl daiktavardžiai sutampa mažiausiai, yra ta, kad daugelis jų yra tikriniai vardai ir su siužetu susiję elementai. Be to, žodynų profiliavimas naudojant įrankį "VP Profiler" rodo, kad visos septynios tirtos knygos remiasi dažniausiai vartojamais žodžiais. K1 ir K2 dažnumo lygių žodžiai sudaro daugiau nei 94 procentus kiekvieno teksto. Tai rodo, kad knygos yra iš esmės prieinamos vidutinio lygio besimokantiems ir tinka skaitymo gebėjimų ugdymui. Be to, „GSE Teacher Toolkit" tekstų analizės įrankis patvirtina, kad dauguma knygų atitinka B1+ lygį ir išlieka stabilios pagal raiškos sudėtingumą. Tuo pačiu metu rezultatai rodo nedidelius, bet svarbius skirtumus tarp knygų leksikos įvairovės ir apdorojimo sudėtingumo. Galiausiai, kolokacijų analizė buvo skirta nustatyti dažnus "veiksmažodžio + daiktavardžio" ir "būdvardžio + daiktavardžio" junginius. Ji parodė, kad adaptuoti skaitiniai suteikia besimokantiesiems galimybę pratintis prie natūraliai raiškai būdingų kolokacijų, dažnai vartojamų gimtakalbių produkuojamoje anglų kalboje. Apskritai, rezultatai rodo, kad Kazachstane naudojami adaptuoti skaitiniai yra naudingi žodynui mokytis ir skaitymo gebėjimams ugdyti

Abstract

This study investigates the lexical characteristics of seven English graded readers used in Kazakhstan. It evaluates how effectively graded readers support vocabulary learning at the intermediate level. The study is based on a corpus-driven approach and analyses a self-compiled corpus of graded readers. The corpus consists of 246,969 tokens taken from seven graded readers The main aim is to examine whether the vocabulary in these books reflects general English. To answer this question, the study applies four analytical procedures: single-word analysis, collocation analysis, vocabulary profiling and readability analysis. The single word analysis focuses on top fifty frequent nouns, verbs, and adjectives and compares them with the BNC2014- fiction. The analysis of single words helps to evaluate how closely the graded readers reflect general English usage. The results demonstrate that verbs have the strongest overlap with general English, while nouns show the weakest overlap. The reason why nouns have the weakest overlap is that many of them are proper names and story-specific items. Moreover, vocabulary profiling with the *VP Profiler* shows that all seven graded readers rely heavily on high-frequency vocabulary. K1 and K2 words cover more than 94 percent of each text. This suggests that the books are generally accessible for intermediate level learners and suitable for fluent reading. Furthermore, the GSE Teacher Toolkit confirms that most of the books are aligned with the B1+ level and remain stable in readability. At the same time, the results show small but important differences in lexical density and processing difficulty across books. Finally, the collocation analysis examines frequent verb+ noun and adjective + noun combinations and shows that the graded readers provide learners with meaningful natural combinations that are used in general English. Overall, the findings show that graded readers used in Kazakhstan are useful for supporting vocabulary learning and reading fluency.

Acknowledgments

I would like to express my sincere gratitude to Vilnius University and the Faculty of Philology. My time within the English Studies (Linguistics) program has been rewarding, providing me with academic knowledge and environment necessary to complete this work.

I am especially grateful for my supervisor Rita Juknevičienė for her invaluable guidance, patience, and expertise throughout the research process.

I also want to thank Ilyas and Roman. Your mental support and friendship helped me when the finish line felt far away. I could not have asked for better companions during this journey.

Finally, my deepest thanks go to my family. Your belief in me has been my greatest source of strength. A special mention belongs to my grandmother, who is no longer with us. I hope you are up there looking down on us with a smile on your face, seeing your Nonash getting one step closer to everything he promised. I carry your memory with me every step.

Table of contents

1. Introduction

In the context of English as a Foreign Language (EFL) education, vocabulary knowledge plays a crucial role in learners' ability to communicate effectively and comprehend authentic language input. Textbooks continue to serve as the primary instructional resource for English language learning, particularly in primary and secondary schools. Therefore, the lexical content of the textbooks largely determines the quantity and quality of vocabulary exposure available to learners. While communicative competence depends on a range of linguistic skills, vocabulary remains a fundamental component of successful language acquisition. Learners' lexical knowledge not only facilitates reading comprehension and writing accuracy, but it might also support fluency in speaking. Consequently, an investigation into the vocabulary presented in EFL textbooks might offer valuable insights into how well these materials support learners' lexical development.

To better understand the focus of the current study, it might be important to review earlier studies on vocabulary in EFL materials conducted in different educational contexts. Previous study in the context of South Korea (Shin & Chon, 2011) have revealed considerable variation in the vocabulary load, frequency distribution, and selection rationale of textbooks. The study frequently highlights a lack of consistency in word choice, with some textbooks including large numbers of low-frequency or academic words that may not correspond to learners' communicative needs. A similar concern has also been addressed through analyses of children's literature and school textbooks. Foster and Mackie (2013) examine the vocabulary frequency coverage of Dr. Seuss's books and compared them with the VP-Kids Corpus and the British National Corpus (BNC) to assess their appropriateness for EFL learners. Their findings reveal that 86 percent and 84 percent of the words in the Seuss corpus are among the 1000 most frequent words in the VP-Kids corpus and the BNC, respectively. The results suggest that the books are representative of both children's language and general English usage. Moreover, their comparison of the most frequent lexical verbs, adjectives, and nouns in Dr. Seuss's writing to those in a children's literature corpus demonstrate strong correspondence, particularly for verbs and adjectives.

Another relevant investigation is that of Konstantakis and Alexiou (2012), who analyse the vocabulary in five EFL textbooks used in Greece during the first two years of primary education. Their analysis is based on a comparison with the 2000-word list from the BNC and show that the books covered between 74 percent and 85 percent of the most frequent words. They conclude that such vocabulary coverage allows only the most basic communicative ability. Konstantakis and Alexiou (2012) further observe that the vocabulary across the Greek books is highly varied in both selection and length, indicating a lack of systematic planning in lexical input.

Expanding on these international findings, Norberg and Nordlund (2018) conducted a corpus-based analysis of lexical input in seven EFL textbooks used in Swedish primary schools for years 3 and 4 (students aged 9-10). Their research is completed in the Swedish Young Learner Corpus (SWYLC), comprising 34,380 running words extracted from textbooks. The analysis focused on content words nouns, verbs, and adjectives which represent about one-third of the total corpus (12,714 words). Using concordance software such as MonoConc Pro and WordSmith Tools, the authors calculate standardized type-token ratios (STTR) to assess lexical diversity and recycling across books. To evaluate the extent to which the vocabulary reflects common language use, the SWYLC is compared with two reference corpora: the New General Service List (NGSL), and the VP-Kids Corpus. The comparisons show that while the Swedish textbooks include a substantial amount of high-frequency vocabulary, there is a notable variation in lexical density, word recycling, and selection consistency among the books. These results align with earlier findings by Konstantakis and Alexiou (2012) and Foster and Mackie (2013), supporting the observation that EFL materials for young learners often lack standardization in lexical choice and frequency distribution. Norberg and Nordlund conclude that although the textbooks generally correspond to everyday language use, the absence of a systematic rationale for vocabulary selection might limit learners' sustained exposure to core lexis essential for effective language acquisition.

Existing analyses have primarily been conducted in other educational contexts and other educational materials, leaving a gap in research concerning the lexical appropriateness of EFL graded readers used in Kazakhstan. This study seeks to address this gap by examining whether the vocabulary presented in English graded readers used in Kazakhstan reflects frequency and coverage patterns that effectively support vocabulary learning. Moreover, it aims to explore whether these textbooks share a common rationale in terms of word selection and whether the included lexical items correspond to those found in everyday English usage.

The purpose of this study is to investigate the lexical complexity of English as a Foreign Language (EFL) Graded Readers used in Kazakhstan by employing a corpus-driven approach. The research focuses on identifying the frequency, range, and distribution of vocabulary across selected textbooks, as well as assessing the extent to which the vocabulary aligns with high-frequency lexical items in established reference corpora. By doing so, the study aims to determine the adequacy of lexical input provided to learners and to evaluate the consistency of vocabulary presentation across graded readers of comparable levels.

To achieve this purpose, the study is going to answer the following research questions:

1. To what extent do EFL graded readers used in Kazakhstan reflect vocabulary that corresponds to high-frequency words found in general English corpora?
2. How consistent is the lexical distribution across books at the same educational level?

3. Is there evidence of a shared rationale or framework guiding vocabulary selection across different books?

The main objectives of this research are:

1. To compile a representative corpus from EFL graded readers used in Kazakhstan.
2. To analyse the lexical frequency, range, and type-token ratio of vocabulary in these books.
3. To compare the vocabulary profiles with established reference corpora to assess lexical coverage and appropriateness.
4. To identify patterns or inconsistencies in vocabulary selection across books of the same educational level.

The thesis consists of the following parts: an introduction, a theoretical and literature review part, an analytical part, and conclusions. The theoretical part provides an overview of key concepts related to vocabulary acquisition, lexical frequency, and corpus-based textbook analysis. As well as it provides detailed description of earlier studies related to this research. The analytical part presents the findings of the corpus analysis, including statistical and interpretive results. The concluding part summarizes the main findings, discusses their pedagogical implications. By filling the existing research gap, this study aims to contribute to evidence-based improvement of EFL materials and support the alignment of Kazakhstani English language education with international standards such as the Common European Framework of Reference for Languages (henceforth: CEFR, Council of Europe 2001, 2020).

2. Literature Review

In this section, a few important theoretical concepts, necessary for the analysis of the data, are presented. Subsection 2.1 provides a comprehensive definition of "vocabulary"; Subsection 2.2 provides information about the lexical load required for reading. In Subsection 2.3, the Common European Framework of Reference for Languages (CEFR) is introduced as a standardized benchmark for proficiency levels and "can-do" descriptors. Finally, Subsection 2.4 provides a detailed discussion of "graded readers", exploring how these materials are used for controlled input and multimodal support to facilitate second language acquisition.

2.1.The definition of the term vocabulary

Vocabulary plays a major role in the English language learning process. Long and Richards (2007) consider vocabulary as "the core component of all of the language skills" (Long and Richards 2007: 12) and it plays major role "in the lives of all language users, since it is one of the major predictors of school performance, and successful learning and use of new vocabulary is key to membership of many social and professional roles"(Long and Richards 2007: 12).Thornbury also mentions that "[i]f you spend most of your time studying grammar, your English will not improve very much. You will see most improvement if you learn more words and expressions" (Thornbury 2002: 114). The quote

by Long and Richards (2007) explains that vocabulary is interconnected with all language skills, including listening, speaking, reading, and writing. Improvement of vocabulary leads to the improvement of other skills that are mentioned earlier. These quotes prove that vocabulary is very important in language learning process, but what is the meaning of word 'vocabulary'. According to Ur (1991),

> Vocabulary can be defined, roughly, as the words we teach in the foreign language. However, a new item of vocabulary may be more than just a single word: for example, post office, and mother-in- law, which are made up of two or three words but express a single idea. A useful convention is to cover all such cases by talking about vocabulary items rather than words (Ur 1991: 60).

Ur (1991) says that vocabulary is the combination of words, and by mentioning "vocabulary items" he emphasizes knowing the spoken form, written form, the meanings, the collocations of the word. Basically, when a person knows the word, it means that s/he understands how words function in different situations. According to a Scott (2005), "one factor that contributes to the complexity of studying word knowledge is the understanding of what it means to know a word. Knowing a word can range from being able to supply a definition to having a vague understanding of its semantic field" (Scott 2005: 70). From this quote it can be understood that studying the word knowledge is complicated, because knowing a word encompasses a range of knowledge levels. For example, at a basic level, knowing a word might stand for providing a dictionary definition, but the definition alone does not show the complicated nature of knowing a word. According to Thornbury (2002), "at the most basic level, knowing a word involves knowing: its form, and its meaning" (Thornbury 2002: 15). Knowing how the word is written and knowing the meaning of a word is the basic level of knowing a word but knowing the meaning of a word does not mean just providing a definition of a word. According to Thornbury (2002), "knowing the meaning of a word is not just knowing its dictionary meaning (or meanings)- it also means knowing the words commonly associated with it (its collocations) as well as its connotations, including its register and its cultural accretions" (Thornbury 2002: 15). A deeper understanding of a word involves understanding a word's connotations, understanding the grammatical behaviour of a word, and knowing the word's collocations. For example, the word 'take' usually collocates with 'a shower', 'a risk', 'an exam'. Knowing these collocations helps people to understand what it means to know a word. This understanding of vocabulary can be relevant to the study of fiction, because literary texts provide context where words appear in diverse settings. Unlike technical registers, fiction might offer EFL learners' exposure to the words embedded in narrative contexts. By reading fiction, learners can understand how vocabulary functions within a complex narrative structure.

2.2 How much lexis is necessary for reading comprehension

One of the important questions in second language acquisition is how much vocabulary a learner needs to know to understand written texts. Nation and Waring approach this issue by asking "how many words are needed to do the things that a language user needs to do" (Nation & Waring 1997). Their perspective might be relevant for reading comprehension because it focuses on functional vocabulary use rather than total vocabulary size. This means that successful reading depends on knowing enough words to process a text effectively. Although vocabulary is important, it does not operate separately from other aspects of language ability. It is noted by Nation and Waring that "vocabulary knowledge is only one component of language skills such as reading and speaking" (Nation & Waring 1997). This statement explains that successful comprehension involves the combination of several competences. It includes grammatical knowledge and the ability to process discourse. In other words, knowing individual words is not enough if learners cannot relate them to each other within a meaningful structure. However, vocabulary remains an important factor. Without sufficient lexical knowledge, learners might fail to understand even the basic meaning of a text.

An important term in the discussion of the minimum of amount of vocabulary for reading comprehension is text coverage. This term "refers to the percentage of running words in the text known by the readers" (Nation & Waring 1997). This supports a frequency-based approach to vocabulary learning, where learners focus first on the most common words in the language. However, knowing high-frequency words alone is not enough for full comprehension. Nation and Waring mention that "knowing about 2000-word families gives near to 80% coverage of written text" (Nation and Waring 1997). At first, 80 percent coverage might seem like a good number, but in practice it means that a significant number of words remain unknown. Nation and Waring clarify that at this level, "1 word in every 5 … [is] unknown" (Nation & Waring 1997), which might create serious difficulties for comprehension. This finding is supported by earlier research. For instance, Laufer argues that "a minimum of 95% coverage is needed for adequate comprehension" (Laufer 1989, cited in Nation & Waring 1997). This supports the idea that readers must know almost all the words in a text to understand it without any problems. Moreover, a similar conclusion is reached by Hirsh and Nation. They suggest that "95% coverage represents a reasonable level for unassisted reading" (Hirsh & Nation 1992, cited in Nation & Waring 1997), which shows a minimum ability for independent reading. The idea of independent reading is important in this context, because it means that learners read without any help of dictionaries or teachers. For independent readers, reaching a high level of lexical coverage is very important.

Additionally, the importance of the vocabulary is shown in research of inferencing. Nation and Waring note that lower coverage "is not sufficient to allow reasonably successful guessing of the meaning of the unknown words" (Nation & Waring 1997), which suggests that guessing from context

becomes unreliable when too many words are unfamiliar. This idea is supported by Na Liu and Paul Nation, where they mention that learners have trouble inferring meaning under such conditions (Liu & Nation 1985, cited in Nation & Waring 1997). Also, Nagy and Anderson show that comprehension depends heavily on the proportion of known vocabulary in a text (Nagy & Anderson 1984, cited in Nation & Waring 1997). Taken together, these findings show that inferencing cannot be used as a strategy to overcome limited vocabulary. The inferencing strategy might be effective only when learners already possess a sufficient lexical base.

Nation and Waring therefore propose a higher threshold for effective reading comprehension. They state that "at least 95% coverage is needed" and that this level "is sufficient to allow reasonable comprehension of a text" (Nation & Waring 1997). The provided threshold is widely accepted in vocabulary research, and it is a good guideline for both teaching material design and teaching. In addition to this, Nation and Waring suggest that a vocabulary size of "3000-to-5000-word families" (Nation & Waring 1997) is necessary to achieve this level of coverage in many texts. This range can be a realistic and achievable goal for many language learners.

The relationship between vocabulary and text type is another important aspect. Nation and Waring note that "with narrowly focused areas […] a much smaller vocabulary is needed" (Nation & Waring 1997). This shows that texts with limited topics and repeated vocabulary might be easier to understand. These types of texts allow learners to read the words multiple times, which can support both comprehension and retention.

Another key aspect is the ability to learn new vocabulary through reading. Nation and Waring explain that "if one does not know enough of the words on a page […] one cannot easily learn from context" (Nation & Waring 1997). This shows the importance of reaching a certain lexical threshold before reading can become an effective language learning tool. Without enough vocabulary knowledge, learners may struggle to understand the text and fail to learn or acquire new words.

This idea is closely related to the role of extensive reading in language learning. Nation and Waring note that "extensive reading is a good way to enhance word knowledge and get a lot of exposure to the most frequent and useful words" (Nation & Waring 1997). Through repeated exposure, learners can strengthen their understanding of vocabulary and develop good fluency in reading. This process is heavily dependent on reading texts that are at an appropriate level of difficulty.

To sum up, the research by Nation and Waring, together with studies by Laufer, Hirsh, Liu and Nagy and Anderson, demonstrate that vocabulary knowledge plays an important role in reading comprehension. While high-frequency words provide a good foundation, learners must reach a threshold of approximately 95 percent lexical coverage to understand texts effectively and to expand

their vocabulary or learn from them. These findings can be a good theoretical basis for analysing reading materials in terms of their lexical demands and their suitability for language learners.

These findings can be relevant for the present study because they provide a theoretical basis for analysing reading materials in terms of their lexical demands and their suitability for language learners. To be specific, they show that the problem is not only how many words occur in a text, but also whether the proportion of familiar vocabulary is high enough to support comprehension without constant interruption.

Graded readers are specifically designed to provide low-level learners with a controlled lexical environment. By adapting texts to specific proficiency levels, these materials might serve as a bridge that allows learners to meet enough known vocabulary to support unassisted reading and acquisition. Evaluating how these books manage lexical coverage is a central focus of this study.

2.2.The Common European Framework of Reference for Languages

The Common European Framework of Reference (Council of Europe 2001, 2020) provides an essential basis for understanding how language proficiency is structured and described in contemporary language education. According to the framework, it "provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe 2001: 1). This statement is important for the present study, because it establishes that teaching materials, which include graded readers are expected to be designed with reference to shared standards. An important feature of the CEFR is its focus on language use rather than isolated knowledge. It is stated that it "describes in a comprehensive way what language learners have to learn to do in order to use a language for communication" (Council of Europe 2001: 1). This shows the action-oriented nature of the CEFR, where language ability is defined in terms of communicative outcomes. In the context of the graded readers, it might suggest that the texts assigned to a particular level should support learners in achieving specific communicative goals appropriate to that level. Another important aspect of the CEFR is related to its system of proficiency levels. It is stated that it "defines levels of proficiency which allow learners' progress to be measured at each stage of learning and on a life-long basis" (Council of Europe 2001: 1). The levels provide a structured way of organizing language learning and are widely used in educational materials. The graded readers also have level labels, which means that the levels are often aligned with CEFR categories. This makes the framework relevant for evaluating whether such materials correspond to expected proficiency levels.

In Council of Europe (2001), vocabulary range grows from A1 "has a basic vocabulary repertoire of isolated words […]" (Council of Europe 2001: 111) to terms at B2 "a good range of vocabulary for matters connected to his/her field and most general topics" (Council of Europe 2001:

111). After it reaches unlimited scope which includes “a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms” (Council of Europe 2001: 111) at C1-C2. Reading descriptors cover “very short, simple texts” (Council of Europe 2020: 54) at A1, and it goes till the C2 where they “can understand virtually all types of texts including abstract, structurally complex, or highly colloquial literary and non-literary writings” (Council of Europe 2020: 54). This progression in vocabulary and text types might be important for this study, because it provides the benchmark to evaluate whether graded readers scale their lexical content appropriately across level to support learners’ reading comprehension.

On the other hand, the CEFR acknowledges certain limitations in its approach. It is stated that “the taxonomic nature of the Framework inevitably means trying to handle the great complexity of human language by breaking language competence down into separate components” (Council of Europe 2001: 1). This suggests that level descriptors can be useful for organization, but they may not fully capture the complexity of real language use.

Furthermore, the CEFR mentions the importance of clarity and comparability in language education. The principle of transparency is important for textbook and material analysis, because the level of a learning resource should be clearly defined and justified. As it is stated “by providing a common basis for the explicit description of objectives, content and methods, the Framework will enhance the transparency of courses, syllabuses and qualifications” (Council of Europe 2001: 1).

Another important component of the CEFR is the use of so called “can do” descriptors, which operationalize language proficiency levels in practical terms. These descriptors reflect the frameworks’ action-oriented approach. As it was mentioned earlier, the CEFR comprehensively describe what learners need to learn to use language for communication (Council of Europe 2001: 1). In this sense, “can do” descriptors translate levels into observable abilities. For instance, it includes understanding simple texts or identifying main ideas in passages. This is important for the analysis of adapted reading texts, because the materials typically correspond to CEFR levels and are intended to help learners to achieve the communicative outcomes. The “can-do” descriptors provide a useful reference point for evaluating whether the linguistic content and difficulty of graded readers correspond to the skills expected at each level of proficiency.

To sum up, the CEFR is a well-organized framework for describing language proficiency. It is relevant for the analysis of graded readers because of its emphasis on levels, communicative ability of learners, and transparency. Also, its acknowledgment of the complexity of language shows the necessity to investigate how linguistic features are distributed across different proficiency levels.

2.3.Features of Graded Readers

The graded readers are pedagogical texts that have “simplif[ied] grammar, vocabulary and plot” (Waring 2018). According to Waring (2018), one of the main functions of graded readers is “to create a series of stepping stones for foreign language learners to eventually read unsimplified materials”. This definition is relevant to the current research because it highlights the level differentiation. The level differentiation can be a key aspect in the analysis of graded reader series. Waring (2018) further explains that graded readers are written “at different difficulty levels” by “simplifying the grammar, vocabulary and plot as well as adding illustration to ensure adequate comprehension at that level”. This suggest that the linguistic complexity of graded readers increases across levels, which allows learners to progress steadily. At the early stages, learners are exposed to texts with “a very limited vocabulary” and simple grammatical constructions, whereas higher levels contain more advanced vocabulary and grammatical constructions (Waring 2018). Such progression might be important for corpus-driven analysis because it raises the question of how lexical difficulty develops systematically across levels.

Furthermore, visual and textual support is another characteristic that is connected to lexical analysis. Waring (2018) notes that graded readers often include “illustrations to ensure adequate comprehension” (Waring 2018). This means that vocabulary is not presented only through text but is also supported visually, especially at lower levels where learners might have limited lexical knowledge. The use of illustrations might reduce the lexical burden of the text and help learners to understand unfamiliar things through contextual support. It might also mean that the vocabulary load in text itself might not fully represent the overall input the learners can get, which is an important consideration when interpreting corpus results.

In addition to their linguistic structure, graded readers can also serve different pedagogical purposes. According to Waring (2018), “graded readers can be used in several different ways”. The first approach is related to extensive reading, which Waring (2018) describes as “to practise the skill of fast fluent reading”. This means that learners mainly focus on understanding the story and developing reading fluency rather that analysing language in detail. The second approach is intensive reading, where graded readers are used “in a language and focus on form approach to reading activities” (Waring 2018). This suggest that in intensive reading put much emphasis on lexical and grammatical items.

Moreover, the descriptions of graded reader emphasize the intended educational role of graded readers in supporting gradual language development. However, these descriptions do not provide empirical evidence regarding how vocabulary is selected, distributed across different levels. Publishers often claim that graded readers support vocabulary learning and reading fluency, but there is limited evidence showing whether lexical progression truly corresponds to increasing learner

proficiency. There is a difference between what these books are supposed to do and what they do. Even though publishers claim that these books support students learn vocabulary, the evidence or data to prove how they choose words for each level is not provided. It makes the corpus-driven analysis of these materials' important analysis, because it helps to see if they really match the learning goals they claim to support.

To sum up, the discussion of graded readers underlines the level differentiation, vocabulary support, and their role in reading-based language development. However, while these features are presented as intended characteristics, the effects of the features are not measured. This creates a gap between description and analysis, which can be addressed through corpus-driven investigation of the lexical content of graded readers.

3. Research of vocabulary used in EFL textbooks.

In recent decades, several corpus-based studies have examined the vocabulary profiles of English textbooks and children's literature in different national contexts. The present section reviews four key studies that are particularly relevant to this research: Shin and Chon (2011) analysis of English textbooks used in the context of the South Korea; Konstantakis and Alexiou (2012) analysis of English textbooks used in context of Greece; and Jordan Foster and Craig Mackie (2013) who analyse children's literature; and Norberg and Nordlund (2018) analysis of English textbooks used in Sweden. These studies provide important insights into vocabulary frequency, lexical coverage, and consistency in EFL materials.

3.1. Shin and Chon (2011): A corpus-based analysis of curriculum-based elementary and secondary English textbooks

One of the important corpus-based investigations into textbook vocabulary is conducted by Shin and Chon (2011), who analyse elementary and secondary English textbooks used in South Korea. They attempt to test the hypothesis whether vocabulary taught in these textbooks is related to the high-frequency words in English. To do this, they compare textbook vocabulary to West's General Service List (GSL), which is recognised as "a list of the most important, foundational words of general English for second language learners" (Browne, Culligan & Phillips 2013), and Coxhead's Academic Word List (AWL). Coxhead's Academic Word List (AWL) is a resource that "contains 570-word families which frequently appear in academic texts, but which are not contained in General Service List (GSL)" (Smith 2026). According to Shin and Chon (2011, as cited in Norberg and Nordlund 2018: 464), "68% of the words in the textbooks were not included in the GSL". This result raises a question about the suitability of the lexical input for learners at early stages of language acquisition. Moreover, Shin and Chon (2011, as cited in Norberg & Nordlund 2018: 464) note that the textbooks contained "a high number of words were academic words.". While academic vocabulary is important

for advanced learners, its presence in early-level textbooks may increase lexical difficulty. The researchers also compare textbook vocabulary with general corpora of English. “The comparison of the vocabulary with the words in the three corpora showed that the textbooks contain a large number of words used infrequently in everyday language production” (Shin & Chon 2011, cited in Norberg & Nordlund 2018: 464). The study highlights the importance of aligning textbook vocabulary with frequency-based principles. The lack of emphasis on high- frequency vocabulary can restrain the process of building communicative competence in learners. Their study offers evidence which shows that textbook vocabulary must be selected with caution based on empirical frequency information.

3.2. Nikolaos Konstantakis and Thomai Alexiou (2012): Vocabulary in Greek young learners' English as a foreign language coursebooks

Another important study focusing on vocabulary in school textbooks is carried out by Konstantakis and Alexiou (2012). They investigate five EFL textbooks used in the first two years of primary education in Greece. Their study aims to determine whether the vocabulary load of these textbooks is sufficient to support meaningful communication. The authors compare textbook vocabulary to the 2000 most frequent words in the British National Corpus (BNC). Their study reveals that the textbooks cover between 74 percent and 85 percent of the most frequent words on the BNC list. According to the researchers, this level of coverage is insufficient for effective communication. They state that such vocabulary size is “insufficient for anything but the most basic form of communication” (Konstantakis & Alexiou 2012: 40). Interestingly, their conclusions differ from those of scholars such as Nation (1997) who emphasize the importance of mastering high-frequency words before introducing less frequent items. Konstantakis and Alexiou (2012) claim that the “Mid- or low-frequency vocabulary, then, seems to be under-represented in course books” (Konstantakis & Alexiou 2012: 37). This suggests that learners might not receive enough lexical variety to expand their vocabulary beyond the most basic level. Additionally, the study finds significant variation in word selection and textbook length. The authors note that “[…] course books vary in the volume and choice of vocabulary they present” (Konstantakis & Alexiou 2012: 42). This lack of standardization may lead to unequal learning outcomes depending on which textbook is used. For instance, words related to animals or school life may not appear among the most frequent 2,000 words in English but are essential in young learners’ contexts. This demonstrates that vocabulary selection should balance frequency principles with pedagogical relevance. Overall, the Greek study demonstrates that textbook vocabulary may either lack sufficient lexical diversity or fail to provide systematic progression. These findings are particularly relevant when investigating whether similar issues exist in our research.

3.3. Jordan Foster and Craig Mackie (2013): Lexical Analysis of the Dr. Seuss Corpus

A different perspective on lexis in children's literature is provided by analysis of the work written by Theodor Geisel and the analysis are provided by Jordan Foster and Craig Mackie. The "ultimate goal of this project is to run a comprehensive analysis of all of Seuss works using a corpus method" (Foster & Mackie 2013: 7). To address the frequency coverage compared to a general English corpus, the study used the Web-Vp BNC-20 to analyse a "26,263 token corpus" (Foster & Mackie 2013: 10). After removing the proper names and inaccuracies, results show that "K1 words accounted for 84.42% coverage, K2 words for 5.71% coverage and off list words accounted for 2.35% coverage of the corpus" (Foster & Mackie 2013: 10). Notably, the research finds that "95.64% coverage [occurred] at the K5 band level while 98% coverage occurred at the K15 band level" (Foster & Mackie 2013: 10). Furthermore, when comparing the results to a children's spoken corpus suing VP-kids v.9, the analysis finds that the first 250-word band which usually represents "75-80% of children use when speaking their native language" provided only "69.70% coverage" (Foster & Mackie 2013: 12) of the Dr. Seuss corpus. The other important focus of the study is the frequency of lexicology invented by Dr. Seuss. The analysis shows that "Seussian imaginative terms such as Oobleck, kerchoo, and bopulous account for 2.63% of the running words in the Dr. Seuss corpus" (Foster & Mackie 2013: 13). The local educators requested "vocabulary lists of unconventional vocabulary to avoid confusion in their students, who might mistake the invented word for conventional vocabulary" (Foster & Mackie 2013: 18).

Moreover, the results also reveal a "striking difference" between different types of Seussian books. (Foster & Mackie 2013: 13). For instance, the early phonic reader Green Eggs and Ham "is covered almost entirely by K1-band words (94.8% coverage […]) has no off list or imaginative words" (Foster & Mackie 2013: 13). Whereas the standard Seussian text The Butter Battle Book (with 81.43% K1 coverage) is found to be "much more representative of the corpus in general in terms of coverage and off-list words" (Foster & Mackie 2013: 13). The study concludes while the books are "lexically rich", they require "great deal of lexical knowledge for ease in comprehension." (Foster & Mackie 2013: 18)

3.4. Cathrine Norberg and Marie Nordlund (2018): A Corpus-based Study of Lexis in L2 English Textbooks

A more recent and methodologically detailed study is conducted by Norberg and Nordlund (2018) in Sweden. Their research focuses on seven English textbooks used in primary school years 3 and 4. The researchers compile a corpus called the Swedish Young Learner Corpus (SWYLC), which consists of 34,380 running words. The textbooks are scanned, converted into digital text format, and carefully cleaned to eliminate scanning errors. The corpus is tagged according to word class using the BNC Basic C5 Tagset, and concordance software such as MonoConc Pro and WordSmith Tools is

used for analysis. The researchers narrow down the scope to focus specifically on lexical verbs, nouns, and adjectives. They make up 12,714 words of the corpus. To measure lexical diversity, Norberg and Nordlund calculate the standardized type-token ratio (STTR). They explain that "the higher the ratio, the more varied and specialized a particular material is in terms of vocabulary" (Norberg & Nordlund 2018: 465). Since the lengths of textbooks is not the same, the STTR allowed for reliable comparison across materials. The researchers also compare the textbook vocabulary with two reference corpora: the New General Service List (NGSL), which is as it was mentioned earlier "a list of the most important, foundational words of general English for second language learners" (Browne, Culligan & Phillips 2013) and the VP-Kids Corpus. VP-Kids Corpus is a "dedicated profiler […] developed by Roessingh & Cobb, 2008 to match transcribed child language against ten 250-word lemma lists generalized from several empirical studies of oral productions" (Compleat Lexical Tutor 2015). In their findings, they note that while many high-frequency words are included, there is considerable variation in vocabulary distribution and recycling across textbooks. They conclude that although the textbooks generally reflected everyday language use, the absence of a shared framework for vocabulary selection might affect lexical consistency. As they state, there appears to be limited consensus regarding the quantity and selection of words introduced (Norberg & Nordlund 2018: 469-470). This observation aligns with earlier findings from Greece and South Korea.

4. Data and Methods

The present study adopts a corpus-driven research design to investigate the lexical characteristics of English as a Foreign Language (EFL) graded readers used in Kazakhstan. A corpus-driven approach is selected because it enables the systematic, empirical, and quantitative analysis of authentic language data. Unlike intuition-based textbook evaluation, corpus linguistics provides objective evidence about frequency, distribution of linguistic items.

Corpus-driven methods allow researchers to examine large amounts of text, making it possible to compare textbook language with reference corpora that represent general language use. In the context of EFL education, such comparisons might be essential in determining whether learners are exposed to vocabulary that reflects authentic, high-frequency language or whether textbooks rely excessively on low-frequency, literary, or context-specific lexical items. Given that the aim of this study is to evaluate the lexical complexity and representativeness of EFL graded readers used in Kazakhstan, a corpus-driven approach might be both theoretically and methodologically justified.

The primary corpus compiled for this study consists of seven English graded readers currently used in the Kazakhstani educational context. The selected titles are as follows: *This Rough Magic* by Mary Stewart, *The Riddle of the Sands* by Erskine Childers, *Do Androids Dream of Electric Sheep?* By Philip K. Dick, and *The Dead of Jericho* by Colin Dexter. Also, Emily Bronte's *Wuthering*

Heights, Anne Tyler's *The Accidental Tourist*, and Jane Austen's *Sense and Sensibility* are used. Each book was digitized and converted into a machine-readable format to allow computational analysis. The resulting corpus represents continuous running text from the textbooks and reflects the lexical input that learners are most likely to encounter. The size of the corpus is 246,969 tokens.

To assess the representativeness of the vocabulary found in the textbook corpus, the results are compared with frequency data from the **British National Corpus (BNC)**. The BNC "is a 100-million-word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written" (Natcorp 2009), and it is widely regarded as a benchmark for general language use. The written part of the BNC includes "extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text" (Natcorp 2009). To analyse the representativeness of the vocabulary found in the textbook corpus, the results are compared with frequency data from the British National Corpus (BNC), more specifically, its recently released second version BNC2014 Fiction. The fiction subcorpus of the BNC2014 was used to better match the register of the graded readers. By comparing the most frequent lexical items in the corpus of graded readers with those in the BNC2014 Fiction, the study evaluates the extent to which vocabulary of graded readers aligns with authentic English usage.

For the linguistic analysis, the corpus of graded readers (henceforth referred to as the GR Corpus) is processed using **LancsBox X** (Brezina, Timperley, & McEnery 2018), a corpus analysis software package developed for advanced quantitative and qualitative linguistic research. LancsBox X has been chosen for several reasons. First, it integrates frequency analysis, concordance tool, and grammatical tagging within a single platform, which allows for a coherent and efficient workflow. Second, it provides reliable part-of-speech (POS) tagging and lemmatisation, enabling accurate classification of lexical items such as nouns, verbs, and adjectives. Third, the software is widely used in corpus linguistic research, which enhances the methodological credibility and replicability of the study. Fourth, this tool provided access to the BNC2014

Three main tools within LancsBox X that are used: *Words Tool; Text Tool; Grammatical Tagging*. First, the Words Tool in LancsBox X is used to generate frequency lists of all words in the corpus. After generating the lists of all words, the words were grouped according to their parts of speech, focusing on nouns, verbs, and adjectives. By identifying the most frequent lexical items, the study shows which words learners are most likely to encounter in the graded readers. The results from the frequency lists then are compared with the BNC2014 Fiction to check whether the graded readers include common and widely used vocabulary. The second tool that is used is the Text Tool. This tool

is used to calculate general statistics of the corpus. It is used to calculate the total number of tokens and the distribution of texts across the seven graded readers, to see whether the number of tokens among textbooks are balanced or not. Additionally, it is used to ensure that the corpus is balanced and representative. Third, Grammatical Tagging is used to classify words according to their part of speech. This tool makes it possible to focus only on content words and exclude function words such as articles and prepositions which was necessary for the analysis.

In addition to single-word analysis, the study also examines collocations. Collocations are important because vocabulary knowledge includes not only knowing single words but also knowing how words combine in natural language. As Nation states, "knowing what the word means in the particular context in which it has just occurred" (Nation 2001: 26). In this study, collocations are understood as "combinations of words that habitually co-occur in texts and corpora" (Brezina 2018: 67) and show a meaningful lexical relationship rather than appearing together by chance. The study focuses on top fifty most frequent verb+ noun and adjective+ noun collocations. For verb+ noun patterns, the verbs *make, take, give* and *do* were selected because they form many common lexical combinations and they are the most frequently used lexical verbs in the GR Corpus. For adjective+ noun collocations, the adjective *good* was selected because it is one of the most frequent adjectives in English and it was expected that it might appear in many combinations. These collocations were identified using grammatical tagging and concordance analysis in LancsBox X. Firstly, the corpus was searched for all instances of the selected words. Then, concordance lines were examined to identify repeated noun combinations. To make the findings more reliable, collocations were selected not only by frequency, but also by Mutual Information (MI) score. According to Brezina, "collocations can be based either on frequency alone or as is more common, on a statistical measure called an association measure" (Brezina 2018: 67). Based on this information, a word combination was treated as a collocation only if it occurred repeatedly in the graded readers corpus and had a statistical association. While doing this, it is important to keep in mind that "there is no one measure which would suit all purposes and research questions. We therefore need to understand how association measures operate to select the one that best highlights the aspects of the collocational relationship we are interested in" (Brezina 2018: 67).

Furthermore, the study also examines vocabulary profiling to evaluate the lexical load of the graded readers. Vocabulary profiling was carried out using the Vocabulary Profiler available through *The Compleat Lexical Tutor's Vocabulary Profiler* (available at https://www.lextutor.ca/vp/eng/) This tool analyses the text and shows how much of the lexical content belongs to high-frequency and lower-frequency vocabulary. This analysis might be useful in EFL research, because it helps to understand how accessible a text may be for learners and whether lexical demands of the material are suitable for the target level. To perform the analysis, a 1,000- word sample was randomly taken from

the middle section of each graded reader. Each sample was then entered into the tool, which grouped the words according to frequency levels in general English.

Readability and lexical density were also analysed using the *GSE Teacher Toolkit* developed by Pearson (available at https://www.english.com/gse/teacher-toolkit/user/textanalyzer). This tool was used to evaluate the readability level of each graded reader sample with different readability measures and to estimate lexical density. The same 1,000- word samples used for vocabulary profiling were entered into the GSE Teacher Toolkit. It allowed the study to compare graded readers not only in terms of vocabulary frequency, but also in terms of overall text difficulty.

5. Findings and Discussion

This section presents the results of the corpus-driven analysis of the seven selected graded readers, followed by a detailed discussion of these findings in relation to English language usage and previous research. The section is divided into five subsections. Section 5.1 provides information about the corpus size and an analysis of the most frequent nouns, verbs, adjectives compared to the BNC2014Fiction. Section 5.2 provides information about Vocabulary Profiles in Graded Readers. Section 5.3 provides information about the Readability Metrics and Lexical Density in Graded Readers via the GSE Toolkit. Section 5.4 examines the collocational behaviour of core high frequency verbs (*make, do, give, take*). Section 5.5 provides information about collocational behaviour of the adjective *good*. Subsection 5.6 uses these results to compare this study with earlier investigations.

5.1. Corpus Composition and Part-of-Speech Distribution

A comparison of token counts across the seven textbooks shows that the overall length of the books is relatively consistent, with token numbers ranging from approximately 30,000 to 38,000 running words. Although, some variation can be observed, the differences are not that big. This relative balance might suggest that the textbooks are designed according to similar curricular expectations regarding content volume and linguistic exposure. From a methodological perspective, this similarity in size is important because it ensures that the results are not affected by the length of the texts. Since the graded readers are comparable in size, any differences can be linked to the publishers' word choice rather than the number of words per reader. Table below demonstrates the number of tokens of each graded reader:

Table 1. The composition of the GR Corpus

Graded Reader	Number of tokens
This Rough Magic	33,630
The Riddle of the Sands	34,563
Do Androids Dream	38,394
The Dead of Jericho	30,605
Wuthering Heights	35,368

The Accidental Tourist	38,331
Sense and Sensibility	36,078

Additionally, corpus consists of 246, 969 tokens. In addition to raw token numbers, the distribution of each graded reader within the corpus can be seen in the figure below:

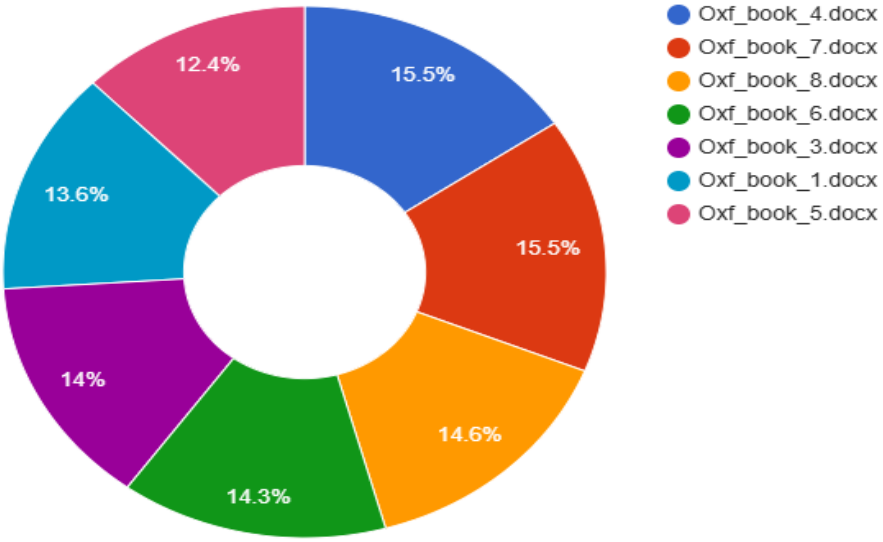

Figure 1. Distribution of books in the GR Corpus (percentage of total tokens)

The pie chart shows the entire corpus of graded readers, with percentages indicating the share of each reader within that total. Each text contributes a relatively small proportion, ranging from 12 percent to 15 percent of the total corpus. Proportional consistency shows that learners are exposed to a similar number of words in all the readers. The number of words might change from level to level, but in level 5 of the levelling system, the length of graded readers is similar.

Moving to an analysis of single words, the noun analysis reveals a markedly lower degree of overlap with the BNC2014Fiction. It is important to note that, the top 50 nouns are taken into consideration, the nouns below top 50 are not used for the analysis. Out of the 50 most frequent nouns in the GR Corpus, only 14 nouns appear among the 50 most frequent nouns in the BNC2014Fiction, which equals to 28 percent. Table below shows the top 50 frequent nouns in the GR Corpus:

Table 2. Absolute Frequency Distribution of Nouns

Noun	**Freq.**	**Noun**	**Freq.**	**Noun**	**Freq.**	**Noun**	**Freq.**
macon	394	*time*	375	*man*	340	*house*	330
elinor	320	*mr*	303	*heathcliff*	296	*rick*	287
day	281	*morse*	276	*I*	263	*edward*	261
davies	257	*door*	242	*sir*	231	*way*	229
room	228	*something*	227	*android*	213	*people*	212
mrs	207	*thing*	189	*catherine*	185	*max*	183

home	181	*night*	179	*face*	177	*muriel*	176
oxford	169	*hand*	166	*police*	161	*life*	161
friend	158	*eye*	157	*godfrey*	157	*charles*	155
nothing	155	*father*	154	*boat*	151	*letter*	150
moment	147	*mother*	147	*edgar*	146	*year*	146
anne	145	*story*	144	*woman*	142	*book*	141
julian	140						

The shared nouns include core lexical items such as *time, man, day, people, way, thing, home, life, year,* and *woman*. These nouns are central to everyday communication and reflect authentic language use. However, a substantial proportion of high-frequency nouns in the textbook corpus consists of proper names and narrative-specific items such as *Macon, Heathcliff, Elinor, Catherine, Edward, Anne*. While such nouns are necessary within story-based materials, their prominence reduces the proportion of general-purpose vocabulary. On the other hand, such proper nouns might increase the complexity of the readers, which is important for the development of reading skills.

Moreover, the strongest alignment with the BNC2014Fiction is observed in the category of verbs. Similarly to the method that was used to analyse nouns, top 50 verbs are at the focus. Out of the most frequent 50 verbs in the graded readers corpus, 35 verbs are also present among the 50 most frequent verbs in the BNC2014Fiction, which equals to 70 percent. The top 50 verbs from graded readers corpus can be seen in Table 3 below:

Table 3. Absolute Frequency Distribution of Verbs

Verb	**Freq.**	**Verb**	**Freq.**	**Verb**	**Freq.**	**Verb**	**Freq.**
be	7969	*have*	2903	*do*	1980	*say*	1585
go	977	*know*	742	*see*	711	*think*	701
look	634	*come*	594	*tell*	580	*get*	543
ask	516	*take*	399	*want*	376	*leave*	368
find	344	*make*	320	*feel*	305	*hear*	257
give	233	*try*	215	*seem*	213	*turn*	194
sit	190	*kill*	188	*talk*	181	*call*	180
put	176	*reply*	170	*happen*	168	*write*	161
cry	161	*use*	160	*move*	158	*begin*	154
stop	153	*read*	151	*like*	151	*keep*	150
live	150	*speak*	149	*walk*	148	*stay*	146
marry	140	*help*	140	*run*	139	*answer*	137
die	131	*show*	131				

The overlapping verbs include core grammatical and lexical verbs *such as be, have, do, say, go, make, get, see, think, take, come, give, want, and use.* These verbs form the backbone of English clause structure and are essential for communicative competence. The high degree of overlap suggests that textbooks provide learners with substantial exposure to high-frequency verbs that are widely used in authentic English.

The analysis of adjectives in the textbook corpus identifies the 50 most frequent adjectival forms. These are then compared with the 50 most frequent adjectives in the BNC2014Fiction. Table 4 below shows top 50 most frequent adjectives from the graded readers corpus:

Table 4. Absolute Frequency Distribution of Adjectives

Adjective	Freq.	Adjective	Freq.	Adjective	Freq.	Adjective	Freq.
good	252	*young*	169	*old*	162	*little*	158
first	141	*sure*	123	*long*	113	*great*	107
happy	107	*wuthering*	106	*bad*	95	*small*	92
right	89	*wrong*	83	*sorry*	80	*afraid*	76
dead	74	*new*	72	*real*	68	*large*	68
important	67	*full*	67	*angry*	65	*true*	64
strong	60	*dark*	57	*poor*	56	*strange*	56
short	56	*cold*	56	*beautiful*	53	*able*	52
german	52	*ill*	52	*big*	50	*difficult*	49
perfect	49	*open*	48	*late*	45		

Out of the 50 most frequent adjectives in the textbook corpus, 28 items are also found among the 50 most frequent adjectives in the BNC2014Fiction. This corresponds to an overlap of 56 percent. The overlapping adjectives include highly frequent and pedagogically central items such as *good, new, old, small, little, important,* and *able*. This suggests that more than half of the adjectival input in the textbooks reflects general English usage, which is beneficial for learners' communicative development. However, the remaining 44 percent of adjectives consist of items that are either lower-frequency (*angry, strange, dead*), context-specific (*wuthering*), or emotive (*sorry, afraid*). While such adjectives are not inherently inappropriate, their high frequency in textbooks may reduce exposure to more broadly applicable descriptive vocabulary.

5.2.Vocabulary Profile in Graded Readers

This section provides an analysis of the vocabulary profiles of seven graded readers. *The Web VP (Vocabulary Profile)* was utilized for the output. The analysis categorises the lexical items into four primary frequency bands: the first thousand most frequent words (K1), the second thousand most

frequent words (K2), Academic Word List items (AWL), and Off-List words. It also evaluates the lexical density of each text to determine the concentration of information. A short extract from each reader with approximately 1000 words were used, and by examining them a clear pattern of lexical distribution emerges which might reveal how these narratives are structured to accommodate intermediate language learners. Table 5 below summarises the coverage of vocabulary metrics for all books:

Table 5. Vocabulary Profiling Metrics of Graded Readers

Text Title	Total Tokens	K1 Words (%)	K2 Words (%)	AWL Words (%)	Off-List Words (%)	Lexical Density
This Rough Magic	994	91.75%	5.33%	0.10%	2.82%	0.45
The Riddle of the Sands	1019	90.97%	5.10%	0.39%	3.53%	0.48
Do Androids Dream	1068	88.86%	5.34%	1.12%	4.68%	0.47
The Dead of Jericho	1038	93.06%	5.30%	0.67%	0.96%	0.47
Wuthering Heights	981	92.25%	5.91%	0.00%	1.83%	0.46
The Accidental Tourist	1008	87.20%	8.04%	0.20%	4.56%	0.51
Sense and Sensibility	928	92.78%	4.85%	0.11%	2.26%	0.48

An examination of the data reveals that the most frequent one thousand words (K1) of English dominate the lexical composition of all seven text. The number of K1 words ranges from a low of 87.20 percent in *The Accidental Tourist* to the highest of 93.06 percent in *The Dead of Jericho*. This reliance on the foundational vocabulary of the English language might be the characteristic of graded materials. This reliance on K1 words might ensure that the primary narration remains accessible to the reader. Moreover, books such as *Sense and Sensibility* and *Wuthering Heights* also demonstrate high K1 concentrations, by having 92.78 percent and 92.25 percent respectively. This might suggest that despite being classical literature, these specific graded readers maintain controlled vocabulary to facilitate reading comprehension.

Furthermore, the K2 band introduces a slight increase in lexical complexity of texts. *The Accidental Tourist* contains the highest percentage of K2 words at 8.04 percent, whereas Sense and Sensibility have the lowest K2 words at 4.85 percent. When evaluating K1 and K2 bands combined, the data shows that all analysed texts exceed a ninety-four percent threshold. The Dead of Jericho reaches a cumulative coverage of 98.36 percent, and *Wuthering Heights* comes after with 98.16 percent. This combined frequency profile might indicate that learners possessing a vocabulary of the

two thousand most common English words would encounter very few unknown terms in these texts. These results can also mean that the materials might support fluent reading, and readers might guess the unfamiliar words by reading the context.

The presence of Academic Word List vocabulary is minimal across the entire dataset, which was very predictable. Narrative fiction generally avoids scholarly, scientific, highly technical discourse. Do Androids Dream contains the highest number of AWL words, but it is very low at just 1.12 percent. One more interesting finding is that *Wuthering Heights* contains absolutely no AWL words within analysed sample.

Furthermore, the Off-list category (proper nouns; highly specific thematic terms and less frequent vocabulary) shows variance across the corpus. *Do Androids Dream* and *The Accidental Tourist* have the highest off-list percentages at 4.68 percent and 4.56 percent. This increase can be attributed to the specific science fiction elements in *Do Androids Dream. The Dead of Jericho* contains 0.96 percent of Off-list vocabulary, which makes it the most lexically contained text focusing on unfamiliar or specialized terminology.

The analysis of lexical density provides insight into the informational load of each text. *The Accidental Tourist* presents the highest lexical density at 0.51, indicating that over half of its text consists of content words rather than grammatical function words. This higher density suggests a more concentrated delivery of information, which might demand a higher cognitive load from the reader. On the contrary, *This Rough Magic* has a lexical density of 0.45, and *Wuthering Heights* has 0.46. These lower figures indicate a higher reliance on grammatical function words within an analysed sample, and it might potentially create a smoother reading experience that may ease the cognitive load on the reader.

To sum up, the vocabulary profile of seven graded readers shows that they heavily rely on the K1 and K2 words. The observed variations in Off-list words and lexical density might highlight the stylistic and thematic differences between the graded readers. Also, the data demonstrates consistency across varied literary genres. It successfully prioritizes foundational vocabulary as well as introduces controlled amount of lower-frequency lexical items.

5.3. Evaluating Readability Metrics and Lexical Density in Graded Readers via the GSE Toolkit

The following part analyses the readability and linguistic complexity of the same graded readers utilizing metrics from the *Pearson GSE Teacher Toolkit*. Similarly to the analysis of Vocabulary Profile, the selected texts maintain a similar length. Similarly to the method used in Vocabulary Profiling, a short extract from each graded reader was used. The extracts range between 934 and 1004 words, which ensures reliable comparison. The table below shows a comparative overview of text

complexity for seven selected graded readers; it contains information about their Global Scale of English (GSE) and CEFR proficiency levels alongside readability indices:

Table 6. Text Complexity of Graded Readers

Title	GSE level	CEFR level	Word Count	ARI	Coleman Liau	Flesch Kincaid	FOG	SMOG
Do Androids Dream by Philip	55-59	B1+	1003	2.5	5.3	3.5	6.1	7.5
The Accidental Tourist by Anne	54-58	B1+	977	3.8	6.5	4.5	6.7	7.9
The Dead of Jericho by Colin	52-56	B1+	991	3.3	6.0	4.2	6.5	7.7
The Riddle of the Sands by Childers	54-58	B1+	1004	3.5	5.5	4.4	7.0	7.9
This Rough Magic by Mary Stewart	54-58	B1+	974	2.4	4.8	3.9	6.4	7.6
Wuthering Heights by Emily Bronte	54-58	B1+	949	2.5	4.7	3.4	5.4	6.4
Sense and Sensibility by Jane Austen	57-61	B2	934	7.2	8.5	7.3	10.4	10.5

To analyse metrics accurately, it is important to understand what these indices measure. First, ARI and Coleman Liau calculate readability based on sentence length and word length. Second, Flesch-Kincaid focuses on sentence length and the average number of syllables per word. Lastly, FOG and SMOG focus on words with three or more syllables.

The main noticeable thing from the table is the CEFR level of the graded readers. Six of the seven graded readers have B1+ level. The level corresponds with the information provided by the publishers where they marked them as stage 5 level graded readers, which is B1-B2 levelled texts. The Global Scale of English (GSE) scores mostly range between 52 and 59. This might indicate a shared grammatical complexity and vocabulary suitable for intermediate level learners. Even though GSE and CEFR levels group these texts together, the readability indices reveal small structural differences that might affect the reading experience.

Inside the B1+ group level, *Wuthering Heights* has the lowest structural difficulty across nearly all measures. This text has the lowest Flesch Kincaid (3.4), FOG (5.4), and SMOG (6.4) scores. Whereas texts like *Do Androids Dream* and *This Rough Magic* have low scores on ARI with 2.5 and 2.4. This suggests their sentence structures are easy to follow. Moreover, a common pattern across all texts is the high number of SMOG index compared to the ARI. SMOG focuses on long words with many syllables, this difference might mean that while authors use short sentences, they still choose advanced vocabulary.

The notable difference can be observed in *Sense and Sensibility*. It is the only book at the B2 level, with higher GSE range of 57-61. The readability measures show why this text is more difficult. The score for ARI equals 7.2, Coleman Liau 8.5, and Flesch Kincaid 7.3 are almost double the scores of other books, which means that the text uses longer sentences and has a higher character count per word than the counterparts. Moreover, the FOG and SMOG equal to 10.4 and 10.5, which means that the number of complex words is higher, and it might require more cognitive effort from the reader. These scores suggest that a reader needs a much higher level of English to understand *Sense and Sensibility*. The reason for this can be, usage of very long sentences with many parts and more difficult words.

To sum up, analysis using the *Pearson GSE Teacher Toolkit* shows that even if different books are placed on the same level, their grammar and vocabulary can be quite different. These readability scores can be important, because they show how a book like *Sense and Sensibility* is much harder for a reader to process than the other books.

5.4.Verb- Noun Collocations with high-frequency verbs *make, do, give and take.*

Collocational behaviour is further examined through the verb *make*. In the GR Corpus, the most frequent objects of *make* include:

Table 7. The most frequent objects of *make*

Graded Readers corpus	**BNC2014 fiction**
Make a mistake	*Make decision*
Make the effort	*Make sense*
Make arrangements	*Make use*
Make a noise	*Make mistake*
Make conversation	*Make way*
Make good use of	*Make difference*
Make the bed	*Make point*
Make some coffee	*Make contribution*
Make some tea	*Make progress*

A comparison shows that some core collocations are shared, such as:

- *make mistake*
- *make way*
- *make difference*
- *make effort*
- *make arrangement*

- *make choice*

This indicates that essential collocational patterns are preserved. However, the graded readers show a stronger presence of everyday domestic and narrative expressions (*make the bed, make coffee, make conversation*), whereas the BNC2014Fiction contains more abstract, academic, and institutional collocations (*make contribution, make provision, make progress, make payment*).

From a pedagogical perspective, graded readers' emphasis on concrete narrative collocations might help learners to build fluency and reading comprehension through accessible storytelling.

Moreover, the collocations with the verb *take* are also examined. The verb *take* reveals the differences between the two corpora. While there is some overlap in expressions, the distribution of collocations suggests a difference in semantic focus. Table 8 below shows collocations that are found both in the GR Corpus and BNC2014Fiction, and collocations that were found only in BNC 2014Fiction or GR Corpus:

Table 8. Collocations with *take*

Category	**Graded Readers**	**BNC2014 fiction**
Shared collocations	Take care, take place, take time, take notice, take interest	-
Only in BNC2014 fiction	-	take action, take responsibility, take advantage
Only in Graded Readers	Take photograph, take revenge, take hand	-

As it can be seen, in the BNC2014Fiction, many collocations are abstract. For example, the collocations: *take responsibility; take action; take advantage*. These expressions are usually used in academic or formal contexts. Whereas, in graded readers corpus includes more concrete and narrative-based collocations, such as *take photograph* and *take revenge*. These words are related to physical actions and part of the storytelling process. This might suggest that learners are exposed to experiential and visual meanings of *take,* rather than more abstract uses.

Furthermore, collocations of the verb *do* are analysed. When analysing the collocation of the verb, the clear overlap of verbs between GR Corpus and BNC2014Fiction is found. Similarly to Table 8, Table 9 below shows collocations that are found both in GR Corpus and BNC2014Fiction, and collocations that were found only in BNC2014Fiction or GR Corpus:

Table 9. Collocations with *do*

Category	**Graded Readers**	**BNC2014 fiction**

Shared collocations	Do something, do anything, do thing, do job, do work, do harm	-
Only in BNC2014 fiction	-	Do research, do business, do justice
Only in Graded Readers	Do test, do repair	-

The overlap between words might show that learners are exposed to core and high-frequency patterns of English. However, the dominance of *something, anything*, and *thing* in graded readers corpus might suggest that the language is generalised. This may indicate that graded readers prioritize simplicity. These nouns might help to understand meaning without needing specific vocabulary. However, this might also reduce lexical precision. Instead of using expressions like *do research*, learners might rely on more vague expressions like *do something*. Another aspect is related to the cognitive load. General nouns are easier to process and remember, which might support comprehension at lower levels. However, the absence of collocations such as *do business* or *do research* might suggest limited exposure to academic and professional language. Overall, the patterns of *do* suggest that graded readers support basic communication, but the support of the development of more specific lexical knowledge is less developed.

Finally, the collocations with *give* were explored. The collocation with *give* shows stronger alignment with general English. The shared collocations might suggest that learners are exposed to important expressions, especially related to the communication and interaction. The examples can be seen in the Table 10.

Table 10. Collocations with *give*

Category	**Graded Readers**	**BNC2014 fiction**
Shared collocations	Give advice, give information, give answer, give reason, give order, give chance, give permission, give money	-
Only in BNC2014 fiction	-	Give evidence, give support, give attention
Only in Graded Readers	Give talk, give love	-

The results might indicate that graded readers place emphasis on words that are important in everyday life. Expressions such as *give advice, give information, give answer* are essential in everyday communication. Their presence might support learners' ability to interact in real-life

situations. On the other hand, the absence of collocations like *give evidence, give support* might suggest that learners are less exposed to formal and institutional discourse. These expressions are important in academic writing and professional communication. Also, the presence of give talk and give love in graded readers corpus might suggest a narrative and interpersonal focus. This reflects the nature of graded readers, which are often story-based rather than informational or argumentative.

Overall, after analysis of the patterns of graded readers, it can be understood that the books provide a balanced but limited input. Graded Readers focus on concrete, story-based collocations that fit their nature as simple literary works. This naturally means less exposure to academic or professional phrases. This design might help learners understand and remember high-frequency words through easy reading and it might help to build a strong base in everyday English.

5.5. Adjective- Noun Collocations: "Good"

The adjective *good* also shows interesting differences. As it was mentioned in methodology part, the reason why adjective good was chosen for the analysis is because it is one of the most frequent adjectives in English and it might appear in many combinations. In the corpus, the following collocates appear:

Table 11. The collocations of *good*

Graded Readers	**BNC2014 fiction**
Good heavens, good luck, good friends, good reason, good idea, good husband, good news, good opinion	Good idea, good way, good deal, good thing, good news, good reason, good friend, good job, good example, good time, good chance, good practice

There is considerable overlap in core collocations such as*: good idea; good reason; good friend; good news.* This suggests that the graded readers successfully include central high-frequency adjective–noun patterns.

However, the BNC2014Fiction demonstrates a broader semantic range, including evaluative and academic uses such as *good practice*, *good example*, and *good chance*. The graded readers, in contrast, favour interpersonal and narrative contexts (good husband, good luck, good heavens). This pattern reinforces the earlier finding: lexical coverage is relatively representative, but functional diversity is more limited.

5.6. Discussion of the Findings and Comparison with Earlier Studies

The findings of this study provide a clear picture of how graded readers compare to previous investigations into English as a Foreign Language (EFL) textbooks and children's literature. This study shows some differences in how vocabulary is selected and organised in different language teaching materials. To begin with, earlier research on school textbooks often found a lack of

organization. According to Norberg and Nordlund (2018), there appears to be limited consensus regarding the quantity and selection of words introduced (Norberg & Nordlund 2018: 469-470). This is quite different from what we notice in graded readers. Seven graded readers with level 5 are very stable in terms of length, these readers stay consistently between 30,000 and 38,000 tokens.

Furthermore, when looking at the most common words in English, the analysis of graded readers shows a strong focus on core vocabulary which is different from what was seen in previous textbook research. Shin and Chon (2011) found that a substantial proportion of the vocabulary falls outside the most frequent word lists in their analysis of textbooks. Similarly, Konstantakis and Alexiou (2012) mentioned that low coverage of frequent words is "insufficient for anything but the most basic form of communication" (Konstantakis and Alexiou 2012: 40). In contrast, the graded readers in this study show a high alignment with foundational English, where K1 coverage reaches up to 93 percent. The results of the graded readers are similar to what Foster and Mackie (2013) found in their research.

Moreover, a comparison of academic and specialised words provides interesting points. It is noted by Shin and Chon (2011) that the analysed textbooks by them often contain a large percentage of academic words, which can increase the difficulty for students. In contrast, the study of graded readers found that academic vocabulary is almost non-existent in these publications. Although certain books lack academic vocabulary, graded readers constitute adapted literary works. Conversely, the textbooks analysed by Shin and Chon are instructional, and therefore likely to contain a higher density of academic words due to their nature. Instead of academic words, the graded readers include many proper names, which can make the vocabulary list seem unique. This aspect is like the Dr. Seuss corpus, where researchers noted that "Seussian imaginative terms […] account for 2.63% of the running words" (Foster & Mackie 2013: 13). Even though these words are necessary for the story, they do not add the same type of stress as technical academic language. These differences are because of the genres', graded readers are not textbooks with lots of instructions and tasks, whereas textbooks analysed in other studies are designed to have instructional language to explain grammatical concepts.

To sum up, the analysis of the graded readers provides interesting insights regarding their effectiveness as learning materials. Firstly, they prioritize the most useful words in English language. Earlier studies on textbooks, such as the one done by Shin and Chon (2011) shows that a substantial proportion of the vocabulary falls outside of the most frequent word lists however the graded readers have the high K1 and K2 coverage. It ensures that learners are exposed to the core vocabulary they need for fluency and enhancement of reading skills. Another interesting thing is that graded readers in this study show a consistent volume of vocabulary that stays between 30,000 and 38,000 tokens at level 5. While researchers like Norberg and Nordlund (2018: 469-470) found that there appears to be limited consensus regarding the quantity and selection of words introduced. The consistency of words

exposed might suggest that publishers control the number of words learners receive at each level compared to the developers of primary school textbooks in other countries.

6. Conclusions

In conclusion, this study examined the lexical characteristics of seven graded readers used in Kazakhstan to evaluate how well they support vocabulary learning at the intermediate level. The findings show that these graded readers provide learners with accessible language input, through their strong reliance on high-frequency vocabulary.

The analysis showed that all seven readers are similar in length and contribute relatively equally to the GR Corpus. This makes the comparison reliable and shows that differences in results are connected mainly to vocabulary choice rather than text size. Among the three lexical categories (nouns, verbs, adjectives), verbs showed the strongest overlap with general English. The high proportion of shared verbs with the BNC2014Fiction suggests that learners are exposed to many core verbs that are central to everyday communication. Also, adjectives show moderate overlap with general English and provide useful descriptive vocabulary. Nouns showed the weakest overlap because many of the most frequent nouns were proper names and story-specific items.

The vocabulary profiling results showed that all seven graded readers rely on the most frequent words in English. K1 words dominate all texts, and the combined K1 and K2 coverage exceeds 94 percent in every graded reader. This suggest that graded readers can support fluent reading with limited lexical difficulty, but it is important to note that some variation was found in analysis of off-list vocabulary and lexical density. Books like *The Dead of Jericho* were more lexically controlled, whereas books like *Do Androids Dream* and *The Accidental Tourist* contained more off-list items. This might mean that books placed at the same graded level do not always provide the same lexical experience.

The GSE Teacher Toolkit results support this conclusion by showing that all seven graded readers are aligned with the B1+ level and remain stable in overall readability. This confirms that the graded readers are generally appropriate for intermediate learners. However, some books are slightly denser and can be more cognitively demanding, while others are easier to process despite sharing the same level label.

Moreover, further research would be interesting to understand the rationale behind vocabulary selection in these materials. It would be valuable to investigate what guides publishers when making decisions concerning vocabulary of graded readers. To achieve this, the study *"The Text Comes First" - Principles Guiding EFL Materials Developers' Vocabulary Content Decisions* by Bergstrom, Norberg, and Nordlund (2021) can be replicated. Their research showed that textbook developers in Sweden often prioritize engaging texts and rely heavily on intuition, rather than empirical vocabulary

research. Applying similar methodological approach to the developers of graded readers would help to make conclusions on decision-making process when it comes to making graded readers.

References

Corpora

BNC Consortium. *British National Corpus.* Oxford: University of Oxford. Available at: https://www.natcorp.ox.ac.uk/corpus/index.xml

Software

Brezina, V., Timperley, M., and McEnery, T. 2018. *LancsBox v. 4.x.* Lancaster: Lancaster University. Available at: http://corpora.lancs.ac.uk/lancsbox/download.php

Cobb, T. *English VocabProfile (Web VP Classic v.4.1).* Montreal: Compleat Lexical Tutor. Available at: www.lextutor.ca/vp/eng/ Accessed on 2026-04-06.

Pearson. *GSE Text Analyzer.* London: Pearson English. Available at: www.english.com/gse/teacher-toolkit/user/textanalyzer. Accessed on 2026-04-06.

Primary Texts

Austen, J. 2008. *Sense and Sensibility* (Oxford Bookworms Library, Stage 5). Adapted by C. West. Oxford: Oxford University Press.

Brontë, E. 2008. *Wuthering Heights* (Oxford Bookworms Library, Stage 5). Adapted by C. West. Oxford: Oxford University Press.

Childers, E. 2008. *The Riddle of the Sands* (Oxford Bookworms Library, Stage 5). Adapted by P. Hawkins. Oxford: Oxford University Press.

Dexter, C. 2008. *The Dead of Jericho* (Oxford Bookworms Library, Stage 5). Adapted by C. West. Oxford: Oxford University Press.

Dick, P. K. 2008. *Do Androids Dream of Electric Sheep?* (Oxford Bookworms Library, Stage 5). Adapted by A. Hopkins and J. Potter. Oxford: Oxford University Press.

Stewart, M. 2008. *This Rough Magic* (Oxford Bookworms Library, Stage 5). Adapted by D. Mowat. Oxford: Oxford University Press.

Tyler, A. 2008. *The Accidental Tourist* (Oxford Bookworms Library, Stage 5). Adapted by J. Bassett. Oxford: Oxford University Press

Works Cited

Bergström, D., Norberg, C., and 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.

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

Browne, C., Culligan, B., and Phillips, J. 2013. *The New General Service List Project.* Available at: https://www.newgeneralservicelist.com/

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume.* Strasbourg: Council of Europe Publishing.

Foster, J., and Mackie, C. 2013. Lexical Analysis of the Dr. Seuss Corpus. *Concordia Working Papers in Applied Linguistics.* Montreal: Copal Publishing.

Hirsh, D., and Nation, P. 1992. What Vocabulary Size Is Needed to Read Unsimplified Texts for Pleasure? *Reading in a Foreign Language*, 8, pp. 689–696.

Konstantakis, N., and Alexiou, T. 2012. Vocabulary in Greek Young Learners' English as a Foreign Language Course Book. *The Language Learning Journal*, 40(1), pp. 35–45.

Laufer, B. 1989. What Percentage of Text-Lexis Is Essential for Comprehension? In Lauren, C., and Nordman, M. (eds.), *Special Language: From Humans Thinking to Thinking Machines.* Clevedon: Multilingual Matters.

Lextutor. *Vocabulary Profiler (VP Kids).* Montreal: Compleat Lexical Tutor. Available at: https://www.lextutor.ca/vp/kids/ Accessed on 2026-04-16.

Liu, N., and Nation, I. S. P. 1985. Factors Affecting Guessing Vocabulary in Context. *RELC Journal*, 16, pp. 33–42.

Long, M. H., and Richards, J. C. 2007. Series Editors' Preface. In Daller, H., Milton, J., and Treffers-Daller, J. (eds.), *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

Nagy, W. E., and Anderson, R. C. 1984. How Many Words Are There in Printed School English? *Reading Research Quarterly*, 19, pp. 304–330.

Nagy, W. E., Herman, P., and Anderson, R. C. 1985. Learning Words from Context. *Reading Research Quarterly*, 20, pp. 233–253.

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

Nation, P., and Waring, R. 1997. Vocabulary Size, Text Coverage and Word Lists. In Schmitt, N., and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, pp. 6–19. Cambridge: Cambridge University Press.

Norberg, C., and Nordlund, M. 2018. *A Corpus-Based Study of Lexis in L2 English Textbooks.* Luleå: Luleå University of Technology.

Scott, J. A. 2005. *Teaching and Learning Vocabulary.* London: Routledge.

Shin, D., and Chon, Y. V. 2011. A Corpus-Based Analysis of Curriculum-Based Elementary and Secondary English Textbooks. *Multimedia-Assisted Language Learning.*

Smith, S. 2026. "Academic Word List (AWL)." *EAP Foundation.* Last modified 21 February 2026. Available at: https://www.eapfoundation.com/vocab/academic/awllists/

Thornbury, S. 2002. *How to Teach Vocabulary.* Harlow: Pearson Education Limited.

Ur, P. 1991. *A Course in Language Teaching.* Cambridge: Cambridge University Press.

Waring, R. 2018. Writing Graded Readers. *Extensive Reading Central*. Available at: https://www.er-central.com/authors/writing-a-graded-reader/writing-graded-readers-rob-waring/

Santrauka

Šiame tyrime nagrinėjamos septynių Kazachstane naudojamų anglų kalbos skaitymo knygelių, pritaikytų skirtingiems mokymosi lygiams, leksinės savybės. Jame vertinama, kaip veiksmingai šios knygos padeda vidutinio lygio anglų kalbos kaip užsienio kalbos besimokantiems asmenims mokytis žodyną. Tyrimas grindžiamas tekstyno analize ir remiasi autoriaus sudarytu tekstyno, kurį sudaro 246 969 žodžiai. Tyrimo tikslas – ištirti adaptuotose skaitymo knygose vartojamą leksiką ir nustatyti, ar šių knygų leksika atspindi bendrąją anglų kalbą. Siekiant šio tikslo, tyrime derinami keturi analizės metodai: atskirų žodžių analizė, kolokacijų analizė, leksikos profiliavimas ir skaitomumo analizė.

Atskirų žodžių analizė sutelkta į dažniausiai pasitaikančius daiktavardžius, veiksmažodžius ir būdvardžius bei jų palyginimą su BNC2014 tekstyno grožinės literatūros patekstyniu. BNC2014 grožinė literatūra pasirinkta kaip bendrosios anglų kalbos pavyzdys. Rezultatai rodo, kad veiksmažodžiai labiausiai sutampa su autentiška anglų kalba ir suteikia besimokantiesiems dažną bei komunikaciniu požiūriu ai naudingą leksiką. Būdvardžiai taip pat rodo vidutinį sutapimą, o daiktavardžiai skiriasi labiausiai. Daiktavardžiai turi mažiausią sutapimą su BNC2014 grožinės literatūros patekstyniu, nes daugelis dažniausiai pasitaikančių daiktavardžių yra tikriniai vardai ir kiti su knygų siužetu susiję vardažodžiai .

Kolokacijų analizė sutelkta į dažnus veiksmažodžio ir daiktavardžio junginius su veiksmažodžiais *make*, *do*, *give*, t*ake*, taip pat būdvardžio *good* ir daiktavardžių junginius. Rezultatai rodo, kad adaptuotos skaitymo knygos suteikia mokiniams galimybių mokytis natūraliai raiškai būdingos kalbos.

Įrankio “VP Profiler” rezultatai rodo, kad visos septynios adaptuotos skaitinių knygos parašytos dažniausia anglų kalbos leksika. Kiekvienoje knygoje K1 ir K2 dažnumo lygių žodžiai sudaro daugiau nei 94 procentus leksikos. Tai rodo, kad adaptuotos knygos yra prieinamos vidutinio lygio besimokantiems ir tinka skaitymo gebėjimų ugdymui. Vis dėlto kai kuriose knygose vartojama ir retesnės leksikos, todėl jos gali būti šiek tiek sudėtingesnės nei kitos. “General Scale of English” platformoje esanti tekstų analizės priemonė patvirtina šią išvadą, parodydamas, kad dauguma knygų atitinka B1+ lygį ir jų sudėtingumas išlieka palyginti stabilus. Apibendrinant, rezultatai rodo, kad Kazachstane naudojamos adaptuotos skaitymo knygos yra naudingos žodynui mokytis ir sklandžiam skaitymui ugdyti.

Summary

This study investigates the lexical characteristics of seven English graded readers used in Kazakhstan. It evaluates how effectively they support vocabulary learning for intermediate EFL learners. The research is based on a corpus-driven approach and uses self-compiled corpus of 246,969 tokens taken from seven graded readers. The aim of the study is to examine vocabulary used in graded readers, and to see whether the vocabulary in these books reflects general English. To achieve the aim of the study, the study combines four analytical procedures: single-word analysis, collocation analysis, vocabulary profiling, and readability analysis.

The single-word analysis focuses on the most frequent nouns, verbs, and adjectives and compares them with the BNC2014Fiction. The BNC2014Fiction is chosen as general English representation. The results show that verbs have the strongest overlap with general English and provide learners with frequent and communicatively useful vocabulary. Adjectives also show moderate overlap, while the nouns show the weakest overlap. The nouns have the lowest overlap, because many of the most frequent nouns are proper name and story-specific items.

The collocation analysis focuses on frequent verb+ noun combinations with *make, do, give, take* as, well as adjective+ noun combinations with *good*. The results show that the graded readers contain natural and repeated collocational patterns.

The VP Profiler results show that all seven graded readers rely heavily on high-frequency vocabulary. In each graded reader, K1 and K2 words account for more than 94 percent of the lexical content. This suggests that the graded readers are generally accessible for intermediate learners and suitable for fluent reading. Moreover, some books contain more off-list vocabulary and higher lexical, which might make them slightly more demanding than others. The GSE Teacher Toolkit supports this finding by showing that most books are aligned with the B1+ level and remain relatively stable in readability. To sum up, the findings show that graded readers used in Kazakhstan are useful for supporting vocabulary learning and reading fluency.