LINGUISTIC COMPLEXITY IN EFL TEACHING MATERIALS:

A GRADE LEVEL COMPARISON OF EFL TEXTBOOKS IN TURKEY

EMİNE TAŞPINAR

BOĞAZİÇİ UNIVERSITY

2022

LINGUISTIC COMPLEXITY IN EFL TEACHING MATERIALS: A GRADE LEVEL COMPARISON OF EFL TEXTBOOKS IN TURKEY

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

English Language Education

by

Emine Taşpınar

Boğaziçi University

2022

Linguistic Complexity in EFL Teaching Materials:

A Grade Level Comparison of EFL Textbooks in Turkey

The thesis of Emine Taşpınar

has been approved by:

Assist. Prof. Şebnem Yalçın _____
(Thesis Advisor)

Prof. Yasemin Bayyurt _____

Prof. Eric Friginal _____
(External Member)

January 2022

DECLARATION OF ORIGINALITY

I, Emine Taşpınar, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature……

Date ……

ABSTRACT

Linguistic Complexity in EFL Teaching Materials:

A Grade Level Comparison of EFL Textbooks in Turkey

The present study aimed to measure text difficulty of the L2 English textbooks used in the state schools in Turkey in terms of their readability and linguistic complexity (i.e., syntactic and lexical complexity). It further compared two locally published textbooks to two internationally published comparable textbooks to check the similarity of the texts in terms of their readability and linguistic complexity. A corpus of 765 texts, more specifically 329 reading texts and 436 listening texts, were extracted from one series of L2 English textbooks currently used in Turkish state schools from Grade 2 to 12 and two internationally published textbooks (i.e., Interchange3 and Passages1). Overall, 3 readability formulas, 12 syntactic complexity measures and 8 measures of lexical complexity were used to analyze the texts through three online computational tools (i.e., Coh-Metrix, L2 Syntactic Complexity Analyzer (L2SCA) and Lexical Complexity Analyzer (LCA)). The results indicated that linguistic complexity of both reading and listening texts showed a gradual increase along with the increase in the grade levels while, with some exceptions, many measures stabilized across adjacent grades or clusters of grade levels. Although the results of the comparisons among locally and internationally published textbooks were mixed, they supported the findings of the comparisons of the textbooks used in state schools. The results of the study may have implications for the design and evaluation of L2 English teaching materials.

ÖZET

İngilizce Yabancı Dil Materyallerinde Dilbigisel Karmaşıklık: Türkiye'deki İngilizce Yabancı Dil Kitaplarının Seviyelerine göre Karşılaştırması

Bu çalışma Türkiye'de devlet okullarında kullanılan İngilizce yabancı dil kitaplarındaki metin zorluğunu okunabilirlik ve dilbigisel karmaşıklık (sözdizimsel ve sözcüksel karmaşıklık) açısından incelemeyi amaçlamaktadır. Buna ek olarak, okunabilirlik ve dilbigisel karmaşıklık açsından benzerliklerini incelemek amacıyla ülkemizde basılan iki yabancı dil kitabını yurtdışında basılan benzer seviyede iki yabancı dil kitabı ile karşılaştırmaktadır. Bu amaçla hâlihazırda Türkiye'de devlet okullarında 2. sınıftan 12. sınıfa kadar kullanılan bir seri İngilizce ders kitabından ve yurtdışında basılan iki ders kitabından (Interchange3 ve Passages1) alınan 329'u okuma metni ve 436'sı dinleme metni olmak üzere toplamda 765 metinden oluşan bir derlem hazırlanmıştır. Metinler üç çevrimiçi araç vasıtasıyla (Coh-Metrix, L2 Syntactic Complexity Analyzer (L2SCA) ve Lexical Complexity Analyzer (LCA)) toplamda 3 okunabilirlik formülü, 12 sözdizimsel karmaşıklık ölçüsü ve 8 sözcüksel karmaşıklık ölçüsü üzerinden incelenmiştir. Sonuçlar ders kitaplarındaki dinleme ve okuma metinlerinde seviye artışına paralel olarak dilbigisel karmaşıklıkta artış olduğunu göstermiştir; ancak, bazı istisnalar haricinde, birçok ölçü birbirini takip eden iki veya ikiden fazla seviyede stabil kalmıştır. Yurtdışında ve yurtiçinde basılan kitapların karşılaştırmasında karışık sonuçlar elde edilmesine rağmen sonuçlar devlet okullarında kullanılan kitapların karşılaştırmalarına ilişkin bulguları desteklemektedir. Çalışma ikinci dil olarak İngilizce öğretim materyallerinin hazırlanma ve değerlendirmesi konusunda önemli sonuçlar sunabilir.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Assist. Prof. Şebnem Yalçın for her continuous guidance, encouragement and support throughout the process. Without her, this thesis wouldn't have been written. I also would like to thank the committee members Prof. Yasemin Bayyurt and Prof. Eric Friginal for their invaluable feedback and suggestions on my thesis.

I am also grateful to Hasan Ayhan Akbaş for his support and understanding during my master's study. I must also thank my friends for the constant motivation they provided during my studies.

I further would like to thank Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) for the financial support they provided for my studies.

Last but not least, my deepest thanks would be to my dear family who has always supported me and believed in me, which made it possible for me to achieve my goals.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION....1

1.1 Background to the study....3

1.2 The significance of the study....4

1.3 The organization of the thesis....5

CHAPTER 2: LITERATURE REVIEW....7

2.1 Assigning grade levels to textbooks.... 7

2.2 Textbook evaluation research....22

2.3 Educational policies in Turkey....29

CHAPTER 3: METHODOLOGY....38

3.1 Research questions....38

3.2 Corpus for the study....40

3.3 Measures and indices....47

3.4 Analysis of the texts....55

3.5 Statistical analysis....55

CHAPTER 4: RESULTS....59

4.1 Results for research question 1....59

4.2 Results for research question 2....76

4.3 Results for research question 3....83

CHAPTER 5: DISCUSSION....93

5.1 Discussion....93

5.2 Conclusion....103

5.3 Implications of the study....104

5.4 Limitations....106

5.5 Suggestions for future research.....107

APPENDIX A: THE TEXTBOOKS IN THE CORPUS.....109

APPENDIX B: EFFECT SIZES (Cohen's *d*) FOR POST-HOC COMPARISONS AMONG GRADE LEVELS.....110

REFERENCES.....116

LIST OF TABLES

Table 1. English Language Curriculum in Turkey.....32

Table 2. Details of the Texts in the Corpus.....47

Table 3. Descriptive Results of Readability Formulas for the Texts in Locally Published Textbooks.....61

Table 4. Descriptive Results of Length-Based Measures of Syntactic Complexity for Reading Texts.....64

Table 5. Descriptive Results of the Remaining L2SCA Measures for the Reading Texts.....66

Table 6. Descriptive Results of Three Syntactic Complexity Measures of Coh-Metrix for Reading Texts.....70

Table 7. Descriptive Results of Lexical Diversity Measures for Reading Texts.....73

Table 8. Descriptive Results of Syntactic Complexity Measures for Listening Texts.....78

Table 9. Descriptive Results of Lexical Diversity Measures for Listening Texts.....81

Table 10. Descriptive Results of Four Measures of Syntactic Complexity for Comparing Reading Texts.....85

Table 11. Descriptive Results of Lexical Sophistication Measures for Comparing Reading Texts.....89

LIST OF FIGURES

Figure 1. A taxonomy for L2 complexity.....16

Figure 2. Means of length-based measures across grade levels.....65

Figure 3. Means of the remaining L2SCA measures across grade levels.....66

Figure 4. Means of three Coh-Metrix syntactic complexity measures across grades.....70

Figure 5. Means of lexical diversity measures of LCA across grade levels.....75

Figure 6. Means of length-based syntactic complexity measures for listening texts across grade levels.....79

Figure 7. Means of lexical diversity measures for listening texts across grade levels.....82

LIST OF APPENDIX TABLES

Table B1. Effect Size (Cohen's *d*) for Post Hoc Comparisons for Readability Formulas among Grade Levels.....110

Table B2. Effect size (Cohen's *d*) for Post Hoc Comparisons for Measures of L2SCA among Grade Levels for Reading Texts.....110

Table B3. Effect size (Cohen's *d*) for Post Hoc Comparisons for Syntactic Complexity Measures of Coh-Metrix for Reading Texts among Grade Levels.....111

Table B4. Effect size (Cohen's *d*) for Post Hoc Comparisons for Lexical Complexity Measures of Coh-Metrix for Reading Texts among Grade Levels.....112

Table B5. Effect size (Cohen's *d*) for Post Hoc Comparisons of Linguistic Complexity Measures for Listening Texts among Grade Levels.....113

Table B6. Effect size (Cohen's *d*) for Post Hoc Comparisons of Syntactic Complexity of Measures of L2SCA for Reading Texts for Comparison.....114

Table B7. Effect size (Cohen's *d*) for Post Hoc Comparisons of Syntactic Complexity Measures of Coh-Metrix for Reading Texts for Comparison.....114

Table B8. Effect size (Cohen's *d*) for Post Hoc Comparisons of Lexical Complexity Measures for Reading Texts for Comparison.....115

Table B9. Effect size (Cohen's *d*) for Post Hoc Comparisons of Lexical Complexity Measures for Listening Texts for Comparison.....115

CHAPTER 1

INTRODUCTION

Exposure to input in the target language is one of the fundamental requirements of second language acquisition (Ellis, 2005). The learners are expected to be exposed to the target language in either naturalistic or instructional environments. In line with the relative importance of input in L2 acquisition, several hypotheses and theories have been proposed, one of the most well-known ones being the Input Hypothesis by Krashen (1985). Krashen (1985) argued that learners should be exposed to sufficient amount of comprehensible input ($i+1$), which can be defined as the input one level ($+1$) beyond the current level (i) of the learner, to be able to acquire a language. He further argued that exposure to comprehensible input and understanding it are the only necessary conditions (along with the low affective filter) for acquiring a language (Krashen, 1985). On a similar vein, he claimed that second language learning classes would benefit the learners in that they would be the source of comprehensible (and grammatical) input, especially for the beginner level learners (Krashen, 1985 as cited in McLaughlin, 1987). Although the Input Hypothesis gained some support, it has also been criticized widely (Izumi, 2003; McLaughlin, 1987; Swain, 1995). Considering the nonnative-like performance of the learners in terms of their production in French-immersion programs, Swain (1995) argued that (forced) output was also necessary in order for the learners to notice the gaps in their interlanguage and the target language as they produce comprehensible output. On a similar vein, Long (1996) further argued that the interaction between the learner/nonnative speaker and the native speaker or an advanced L2 learner would

facilitate language acquisition through the negotiation for meaning, which would provide the learner with comprehensible input.

Although these theories and hypotheses (and many others) differ in the way they define and operationalize the process of second language acquisition and the use of input in that process, they all agree that being exposed to L2 input is a significant factor in L2 acquisition (Ellis, 2005), be it oral or written. When we consider the limited opportunities of EFL learners for interacting with native speakers of the target language, one of the main sources of input would be textbooks, especially in an instructed second language acquisition context (Meunier, 2012).

In many instructed second language acquisition environments, especially in foreign language context, textbooks are still the main input for the learners (Meunier, 2012). Despite the criticism regarding the textbooks and the overdependence on them in second language education, textbooks play a major role in the classroom in that they facilitate language learning in several ways. As suggested by Cortazzi & Jin (1999), textbooks can have various roles in that they can act:

> ... as a teacher, a map, a resource, a trainer and an authority. As a teacher, a textbook gives students relevant information about grammar and vocabulary, as well as English speaking countries and their cultures. As a map, it shows an outline of linguistic and cultural elements as a structured programme and it guides students and teachers to follow the steps taken in previous lessons. A textbook is viewed as a resource as it contains a set of materials and activities available to the teacher from which one can choose. It can also be a trainer for novice teachers who need valuable instructions, support and guidance. As an authority, a textbook is seen as valid, reliable, written by experts and authorized by important publishers or ministries of education. (as cited in Radić-Bojanić & Topalov, 2016, p. 139)

When the various roles and the prevalence of textbooks in second language learning are considered, it is of significance that the textbooks used meet the needs of the target audience and fulfill the objectives of the teaching program. In addition, in order for the materials to be effective in language education, they need to meet

several criteria including but not limited to the inclusion of authentic and contextualized language, engaging components, various genres of both written and spoken language (Crawford, 1995). Therefore, it is necessary that the second language learners are provided with materials, more specifically textbooks, that are suitable to and catered for their needs.

Considering the significance and the role of the textbooks in second language education, the present study aims to see if the textbooks used in state school in Turkey meet the needs of the learners and the criteria on which the preparation of the textbooks is based in terms of one aspect of language learning, more specifically (the complexity of) the input/language presented in the textbooks.

1.1 Background to the study

As the prevalence of English in various aspects of life has increased both internationally and nationally, teaching English as a foreign language has been given primary attention to in Turkey. In order to increase the efficiency of English language education, several policies have been employed since the introduction of English language teaching in Turkey (Kirkgöz, 2007). For the purposes of the present paper, two recent policy changes, which have had a substantial effect on the present curriculum, will be briefly discussed here. The first major policy change was introduced in 1997 as a result of a major curriculum project by the Ministry of Education and the Higher Education Council. Two changes in the new curriculum in 1997 were that English started to be taught at primary school (Grades 4 and 5) and the Communicative Language Teaching (CLT) approach was adopted as opposed to a focus on grammar (Kirkgöz, 2007). A more recent policy change was made in 2012, which introduced the new system of 4 years of compulsory primary school

education, 4 years of middle school education and 4 years of high school education (4+4+4), and the students started to learn English at Grade 2. As a result of both policy changes, the materials and the designs of the lessons were changed, and students started learning English at the age of around 6 (at Grade 2) until graduating from high school with a focus on communicative competence and performance of the learners.

Despite the policy changes and the focus on the improvement of language teaching, there have been several challenges and problems regarding English language education in Turkey (Ayaz, Ozkardas, & Ozturan, 2019; Kirkgöz, 2007; Valizadeh, 2021). Among the problems faced are the ones related to the coursebooks and the teaching materials used in the state schools (Kirkgöz, 2007) as both the teachers and the students are not satisfied with the course materials (Ayaz et al., 2019; Kizildag, 2009). Some of the problems were related to the way the skills were presented in the books, how engaging the students found the activities (Ayaz et al., 2019) and how much the objectives of the textbooks and the overall curriculum met the descriptions of the Common European Framework of Reference for Languages (CEFR) (Yüce & Mirici, 2009). Overall, several studies have shown that the textbooks may not meet the expectations of various stakeholders.

1.2 The significance of the study

Although several studies have shown that there are problems with the materials used in the state schools for English learning, these studies mainly focused on aspects like the type of activities, the amount of culture-specific elements in the textbooks, the amount of match between the objectives of curriculum and the descriptions of the CEFR levels. However, to the best of my knowledge, few studies investigated the

text difficulty and linguistic complexity of the texts presented in the textbooks. Therefore, the present study would offer significant implications for materials design by investigating if there is a gradual increase in the textbooks used in state schools starting from Grade 2 to Grade 12 in terms of the linguistic complexity (i.e., syntactic and lexical complexity) and readability of the texts in the textbooks. It further aims to provide some insights into the level of complexity of the texts by comparing two locally and two globally published comparable textbooks.

The present study also aligns with Turkey's 2023 Education Vision, which covers the educational plans and policies of the Ministry of National Education (MoNE, 2019). It has been reported that the management of education and the decision-making processes will be based on 'data' on both small (i.e., the schools) and large scale (i.e., the ministry) (op cit.). The purpose is to analyze the large-scale data and figure out the problems and possible reasons for and solutions to these problems to enhance the learning process for the students (MoNE, 2019). The present study also focuses on the analysis of the data obtained from the textbooks used in the state schools and aims to offer suggestions for the improvement of one aspect of language teaching in Turkey, namely the material design and the textbooks.

1.3 The organization of the thesis

This chapter provided some preliminary information on the concepts related to the study and gave some background information about the context where the study was conducted. Chapter 2 presents a review of the literature on text difficulty and linguistic complexity (i.e., syntactic and lexical complexity) research. Chapter 3 introduces the methodology followed in the present study to analyze the texts in line with the research questions. Chapter 4 presents the findings of the study in terms of

the readability and linguistic complexity of the texts. Finally, Chapter 5 discusses the main findings of the study related to the previous research, offers some implications for material design and language teaching research based on the findings of the present study, presents the limitations of the study and makes some suggestions for future research.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a review of the literature on the research studies investigating the differences between texts of various levels in terms of the difficulty they pose to the reader. It starts with an introduction to the studies on the traditional ways of measuring "readability" of the texts. Then L2 complexity, approaches to it and more specifically L2 linguistic complexity are presented. Finally, it provides a review of the research on readability and linguistic complexity of the texts in L1 and L2 textbooks used inside and outside Turkey.

2.1 Assigning grade levels to textbooks

Assigning a grade level to textbooks and providing the texts with right level of difficulty to the learners both in L1 and L2 have been a concern and challenge for publishers, curriculum writers and language instructors for many years since the texts that are used by the learners should be within optimal level of difficulty. The texts and the textbooks should not be too difficult or too easy for the learners in order to guarantee the potential benefits of the texts for the learners (McNamara, Graesser, McCarthy, & Cai, 2014). Therefore, predicting and measuring the difficulty, or the ease, of the texts and textbooks have been a major concern in educational research. The most common ways of predicting and measuring text difficulty have been predicting the readability of the texts, more specifically through the use of readability formulas and, relatively recently, measuring the linguistic complexity of the texts using various linguistic complexity measures, especially measure of lexical and syntactic complexity.

2.1.1 Readability

Building on the previous definitions of the term readability, which associated the term with three elements as typology of the text, the interest of the reader in the text and the style of the text (i.e., vocabulary and sentence structure), Dale and Chall (1949) defined the term as the interaction of all these three elements "within a given piece of material that affects the success that a group of readers have with it" (p. 23). In order to assess the readability of a text or a material for certain groups of readers, several methods have been used, and these included the intuitions of the writers or the publishers based on their experience (Klare, 1974) and the results of comprehension tests taken by readers (Klare, 1974), one of the most common ones being 'cloze tests', and wordlists (Dale & Chall, 1948). Since 1920s, hundreds of readability formulas, both traditional and relatively recent ones, have also been designed in order to predict the readability of the texts based on the linguistic features of them (Brown, 1998). Many of these readability formulas predict the readability of the text on the basis of semantic and syntactic difficulty (Fry, 2002). They predict readability mostly through measures of "frequency or familiarity of the words, and the length or syntactic complexity of the sentences" (McNamara, Graesser, & Louwerse, 2012, p. 90). The underlying notion is that as the sentence length increases and the word familiarity decreases, the texts become more difficult to read and comprehend. Many widely used formulas have been designed using these variables and measures.

Flesch Reading Ease (Flesch, 1948) and Flesch-Kincaid Grade Level formulas are among the most popular traditional readability formulas that have been used to measure the readability of a text and assign a grade level to it (Dufty, Graesser, Louwerse, & McNamara, 2006; McNamara et al., 2014). Flesch Reading

Ease assigns a readability score ranging from 0 to 100 to a text as the readability or the ease of the text increases while Flesch-Kincaid Grade Level formula assigns a grade level to the text based on the U.S. grade levels. As the grade level increases, the difficulty of the text increases as well. Both formulas measure readability of a text through the length of the sentence and the length of words (number of syllables for words). Flesch-Kincaid has the formula as follows (McNamara et al., 2014):

$$\text{Grade Level} = .39 \times \text{Words} + 11.8 \times \text{Syllables} - 15.59$$

In the formula, "Words" is the mean number of words of all sentences (sentence length) and "Syllables" is the mean number of syllables per word (word length) (McNamara et al., 2014). The formulas are based on the notion that longer sentences include syntactically complex forms, which place higher cognitive load on the reader, and therefore are more difficult to comprehend. On a similar vein, longer words are difficult to process as they are less frequent in the language and less familiar to the reader; therefore, they make the text or the sentence more difficult to comprehend (Dufty et al., 2006; McNamara et al., 2014).

Considering the weaknesses of the Flesch Reading Formula in terms of "the count of affixes" (Dale & Chall, 1948, p. 12) and "the count of personal references" (p. 14), Dale and Chall (1948) designed a new formula using a word list which consisted of words known by fourth grade-readers. Dale-Chall formula assigns a readability score to the texts based on sentence structure (i.e., average sentence length) and word load, which is measured as "the percentage of words outside the Dale list of 3,000 words" (Dale & Chall, 1948, p. 19). In 1969, Fry developed a readability graph which predicted readability based on average number of sentences per 100 words and average number of syllables per 100 words. The formula correlated highly with the other traditional readability formulas including Flesch and

Dale-Chall (Fry, 1969). Several other formulas have been designed and widely used including SMOG Grading (McLaughlin, 1969), Lexile scores (Stenner, 1996) and Degrees of Reading Power (DRP) (Koslin, Zeno, & Koslin, 1987), all of which mainly predicted readability based on average sentence and word length.

The traditional readability formulas have been widely used by publishers and instructors to assign difficulty scores and grade levels to the texts as they are rather easy to measure. They can provide an overall picture of the difficulty of the texts in accordance with each other as they assign a numeric grade to each text. A number of studies also investigated the reliability and validity of these formulas and showed that they could accurately predict the readability of the texts. In their study investigating the difficulty of the texts through students' oral reading fluency scores, Begeny and Greene (2014) found that Dale-Chall formula predicted the readability of the texts for elementary grade levels in a valid and reliable way. Therefore, it can be claimed that the scale can be useful and effective in especially assigning readability scores to materials in L1 English.

On the other hand, traditional readability formulas have also been criticized for several aspects (Bailin & Graftsein, 2001; Brown, 1998; Davison & Kantor, 1982; Dufty et al., 2006; Klare, 1976; McNamara et al., 2014). One of the main arguments regarding traditional readability formulas was about whether they could accurately predict the difficulty levels of L2 English texts since they were originally designed based on L1 English texts and textbooks. The studies investigating their use in ESL and EFL contexts showed contradicting results. Brown (1998) investigated the difficulty levels of EFL texts using 6 traditional readability formulas (i.e., Flesch Reading Ease, Flesch-Kincaid Index, Fry Grade Level, Gunning Index, Fog Count and Gunning-Fog Index) and compared the results to the mean score of cloze tests

taken by L1 Japanese learners of English. The results showed that the readability formulas did not adequately predict the difficulty of texts in L2 English. Brown later designed the EFL Difficulty Estimate (Brown, 1998) for predicting readability in L2 English texts. However, in another study with Japanese learners of English, Greenfield (1999) used the traditional readability formulas including Flesch Reading and Flesch-Kincaid and the cloze tests based on Bormuth's (1971) corpus, which consisted of 32 academic reading texts of various subjects in L1 English, to investigate the effectiveness of the formulas in predicting L2 English texts' readability (as cited in Greenfield, 2004). He compared the results of the cloze tests taken by L1 and L2 English readers. The results of the study by Greenfield (1999) showed that traditional readability formulas predicted the text difficulty of both L1 and L2 English texts (as cited in Greenfield, 2004). Greenfield (1999) further designed a formula for L2 texts, Miyuzaki EFL Readability Index (Greenfield, 2004). Although both Brown's (1998) EFL Difficulty Estimate formula and Miyuzaki EFL Readability Index (Greenfield, 1999) were designed for L2 English texts, they were similar to the traditional readability formulas in that they both focused on average sentence and word length (Brown, 1998; Greenfield, 2004; Zamanian & Heydari, 2012). The traditional readability formulas were further criticized for focusing only on surface level properties of the texts and it was argued that there was a need for the analysis of the texts on more deeper level language including measures related to cohesion and cognitive processes.

In order to investigate the effectiveness of traditional readability formulas and compare them to the measures of deeper level language, Crossley, Greenfield and McNamara (2008) used two traditional readability formulas and three psychological indices on Coh-Metrix. Coh-Metrix is a computational tool which automatically

analyzes the language of the texts using various measures for cohesion and syntactic complexity, in predicting the readability and text difficulty of texts (Crossley et al., 2008; McNamara et al., 2014). Bormuth's (1971) corpus was used in order to test if cognitive indices of Coh-Metrix were better at predicting the readability of the texts than the traditional readability formulas. A total of 31 texts with a mean length of 269,28 words were used. The cloze test scores of the participants in Greenfield's (1999) study were correlated with the scores from the readability formulas and three variables of Coh-Metrix. Flesch Reading Ease, Flesch-Kincaid Grade Level and Miyazaki EFL Index (Greenfield, 1999) were used as the traditional readability formulas for L1 and L2. Three indices of Coh-Metrix related to the cognitive processes of lexical recognition, syntactic parsing and meaning construction were used in the study. Word frequency information through the CELEX database frequency scores was used for lexical index as the frequency of the words affected the cognitive load of the texts (Crossley et al., 2008). Sentence syntax similarity (i.e., sentence to sentence), which measured the overlapping syntactic constructions both at the phrase level and parts of speech of the sentences, was used as the syntactic index. The measure was used as it is also related to the cognitive processes used by the readers (i.e., decoding of the sentences). As for meaning construction index, content word overlap, which measures the frequency of overlap in the content words between adjacent sentences, was used because word overlap has been found to be related to the comprehension of the texts and the speed of the reader (Crossley et al., 2008). The results showed that Coh-Metrix indices predicted the difficulty and readability of the L2 texts significantly better than the traditional readability formulas (Crossley et al., 2008). The results implied the significance of using

variables and indices related to not only linguistic but also cognitive and psychological dimensions.

The indices used in Crossley et al. (2008) formed the basis of the development of the readability formula Coh-Metrix L2 Reading Index (McNamara et al., 2014). It predicts readability by measuring word overlap, sentence syntax similarity and word frequency. It differs from traditional readability formulas in that it not only measures readability at the sentence and word level, but also includes measures for cohesion between sentences.

Crossley, Allen and McNamara (2011) further investigated the effectiveness of traditional readability formulas and Coh-Metrix L2 Reading Index in assigning a level (i.e., beginner, intermediate and advanced) to a corpus of texts that were intuitively simplified. A corpus of 210,538 words (a total of 300 texts, 100 texts for each level) which was compiled of the texts extracted from an English teaching website and related to news articles and simplified by a team of authors, was used in the study. Two traditional readability formulas, namely Flesch-Kincaid Grade Level and Flesch Reading Ease, and Coh-Metrix L2 Reading Index were used to measure the readability of the texts. The results showed that there were no statistically significant differences between the two traditional readability formulas (i.e., Flesch-Kincaid Grade Level and Flesch Reading Ease) in discriminating different proficiency levels; however, Coh-Metrix L2 Reading Index had significantly more accurate results in assigning a level to the intuitively simplified texts (Crossley et al., 2011).

Although there have been disagreements regarding the use of traditional readability formulas for ESL and EFL, the major criticism towards the formulas was that they were mainly based on the surface structure of a text, or they had a

'unidimensional approach' to text difficulty (McNamara et al., 2014). The sole focus on sentence and word length may be misleading and may not provide a complete picture of why a text is difficult to comprehend. Rather, text difficulty, or readability, involves multiple factors including grammar and syntax, background knowledge of the reader, coherence and their interaction with each other (Bailin & Graftsein, 2001). Coherence of the texts is also argued to affect the readability of the texts (Bailin & Graftsein, 2001; McNamara, Graesser, & Louwerse, 2012).

In order to investigate the effectiveness of the traditional readability formulas and indices of cohesion, Dufty et al. (2006) compared the Flesch-Kincaid Grade Level formula to the automated indices of cohesion on Coh-Metrix. A corpus of texts of up to 5000 words was compiled from 311 textbooks from all grades (i.e., from K to 12). The texts belonged to three genres, narrative, science and social science, and were further categorized into grades as K-3, 4-6, 7-9 and 10-12, which were assigned by the publishers. The texts were later analyzed by the readability formulas and the indices of cohesion on Coh-Metrix. The results showed that Flesch-Kincaid Grade Level formula had a correlation of an R^2 value of .61 with the grade level, while this value increased to .68 when combined with indices of cohesion. 3 out of 8 cohesion indices showed statistically significant contribution to the prediction of the grade level (i.e., Latent Semantic Analysis, incidence of causal verbs and incidence of causal particles). The results showed that either form of measuring readability was insufficient when used solely; however, when cohesion is used with readability formulas, it predicted the readability of the texts better.

Consistent with the results of Crossley et al. (2008), the results of the study highlighted the need for the development and use of measures and indices analyzing the language in the texts at a deeper level as well. Although the traditional readability

formulas predict the overall readability of a text through average sentence and word length to a degree, the indices related to multiple levels of language, or discourse, may provide a better understanding of the factors affecting text difficulty. Considering this, it can be argued that Coh-Metrix L2 Readability Index, although it is still considered a unidimensional formula (McNamara et al., 2014), can provide a relatively more accurate analysis of the readability of L2 texts than the traditional readability formulas as it includes indices and measures for various aspects of text difficulty including cohesion.

Overall, the readability formulas, both recent and traditional ones, can provide a global understanding of the texts regarding their readability or difficulty. The difficulty of the texts can further be analyzed through more fine-grained measures of linguistic complexity on both syntactic and lexical levels, which can provide more in-depth information about the possible difficulty and complexity of the texts across grade levels.

2.1.2 L2 complexity

Complexity has been one of the constructs that have attracted much attention in both L1 and L2 language acquisition research and has been regarded as a valid indication of performance, proficiency and development in language acquisition (Bulté & Housen, 2014). However, there is still not a commonly agreed definition of the term (Bulté & Housen, 2012; Lu, 2017). Several and rather broad definitions have been proposed for the term "complexity" in both L1 and L2 acquisition research. Considering L2 learning, Ellis and Barkhuizen (2005) defined complexity as "the extent to which learners produce elaborated language" (p. 139) while Housen, Kuiken and Vedder (2012) defined the construct as "the ability to use a wide and

varied range of sophisticated structures and vocabulary in L2" (p. 2). Another definition of complexity identified the term as "a property or quality of a phenomenon or entity in terms of (1) the number and the nature of the discrete components that the entity consists of, and (2) the number and the nature of the relationships between the constituent components." (Bulté & Housen, 2012, p. 22). The definitions mainly point to the level of sophistication or elaboration in the various components or constructs of the term complexity. Recently, it has been argued that complexity is a multidimensional construct (Bulté & Housen, 2014; Norris & Ortega, 2009) and Bulté and Housen (2012) further provided a taxonomy of the constructs of complexity, which can be seen in Figure 1. Differentiating between *relative complexity (i.e., difficulty)* and *absolute complexity (i.e., complexity)*, they categorized complexity into three components as linguistic, discourse-interactional and propositional complexity. While the latter two have been relatively new concepts, linguistic complexity has been the focus of research for several decades.

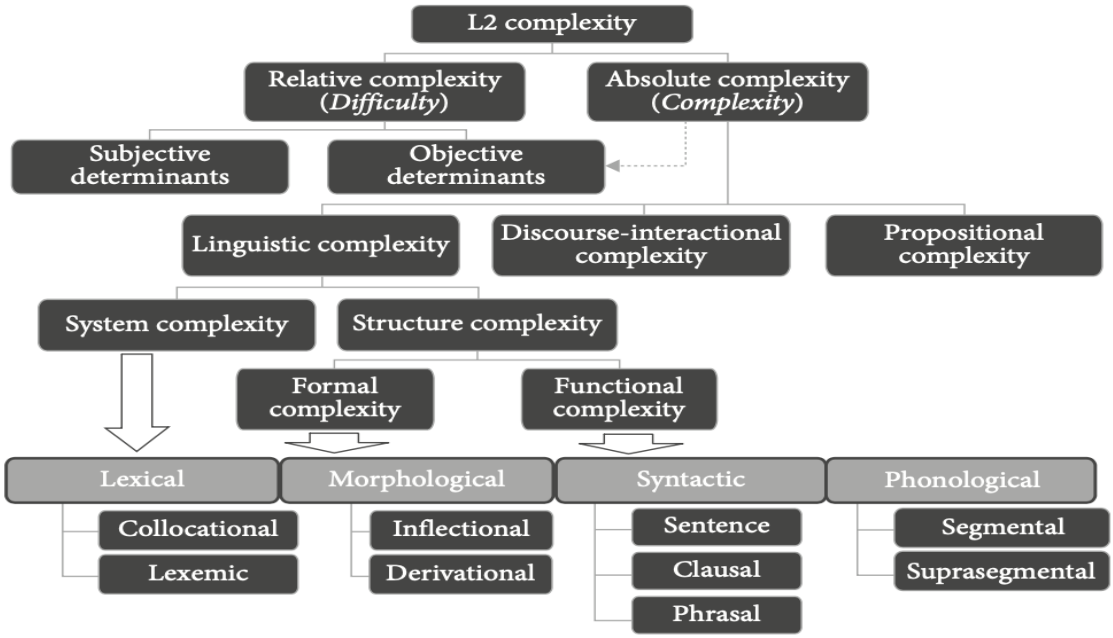

Figure 1. A taxonomy for L2 complexity
Source: [Bulté & Housen, 2012, p. 23]

Regarded as a multilevel phenomenon under the construct of L2 complexity (Green, 2019), linguistic complexity can further be categorized into components as *system* and *structure* complexity considering the way it has been treated in the research (Bulté & Housen, 2012). The former refers to overall number or range of the variety of structures the learners know, or the "global" complexity, while the latter refers to the individual linguistic features that form the knowledge of the learners in the language, or the "local" complexity (op cit.). Various aspects of language including "lexical", "morphological", "syntactic" and phonological" knowledge can be examined either globally or locally.

L2 complexity research has mainly focused on the aspects of lexical and syntactic complexity when assessing the complexity of L2 performance and input. Complexity, combined with accuracy and fluency (i.e., CAF), has been investigated in many studies regarding the proficiency and linguistic development of the L2 learners. The constructs and the measures related to them have been used in especially task-based language research, L2 writing and speaking performance. Several measures and variables of both syntactic and lexical complexity have been produced and used in L2 research to analyze both L2 input and production. For the purposes of the present study, the constructs of syntactic and lexical complexity are further presented.

2.1.2.1 Syntactic complexity

Ortega (2003) defines syntactic complexity as the "range of forms that surface in language production and the degree of sophistication of such forms" (p. 492). It basically refers to the variety of syntactic structures used and the extent of the complexity of the structures used (Lu, 2017). It is a significant construct in L2

learning in that L2 learners' linguistic output and performance are expected to increase in terms of amount, sophistication and appropriate use of various syntactic forms as they become more proficient in the language. Similarly, L2 texts are expected to gradually increase in the variety and amount of syntactically complex structures along with the increase in the proficiency level.

Despite its widespread use in L2 research, there is not an agreement on the operationalization of syntactic complexity. A wide variety of measures have been developed and used to operationalize it (Johnson, 2017; Norris & Ortega, 2009), the most common one being mean length of T-units (MLT), which taps into the overall syntactic complexity. However, many other measures for the subdimensions of the construct have been proposed as well. The main subdimensions include overall or general complexity, which is usually measured through length-based metrics with a multiple clausal unit (e.g., mean length of T-unit, mean length of c-unit), subordination, usually measured by a metric with a clause of subordination or dependent clause (e.g., clauses per T-unit, subordinate clauses per dependent clauses) and subclausal complexity via phrasal elaboration, which is usually measured by mean length of clause (Bulté & Housen, 2012; Ortega, 2003). Conceptualizing syntactic complexity as a multidimensional construct, Ortega (2003) and Bulté and Housen (2012) argued for the inclusion of a variety of measures tapping into different subdimensions of syntactic complexity. While it is important to use a variety of measures for the subdimensions, it is also of significance to avoid redundancy in results as some of the metrics measure the same of dimension of the construct (Ortega, 2003).

Although the measures for syntactic complexity allow the researchers to manually count the number of instances for each measure, the development of

computerized tools now enables the researchers to analyze high number of texts quickly and uniformly using several measures (Crossley & McNamara, 2012). L2 Syntactic Complexity Analyzer (L2SCA, Lu, 2010) is one of the tools that allow the researchers to analyze L2 texts using 14 traditional measures of syntactic complexity, which were selected from the measures covered by Wolfe-Quintero et al. (1998) and Ortega (2003). The measures tap into the indices of the length of production units, overall sentence complexity, the amount of subordination, coordination and degree of phrasal sophistication (Lu, 2010). Coh-Metrix is another computational tool that has several measures and indices for syntactic complexity (e.g., left embeddedness) (McNamara et al., 2014). It provides measures for the amount and density of certain syntactic patterns, word and phrase types (e.g., the incidence of noun phrases), which also inform syntactic complexity. It further provides a measure of syntactic similarity, which measures "the uniformity and consistency of the syntactic constructions in a text" (McNamara et al., 2014, p. 71). More uniform syntactic constructions would lead to a less complex text (op cit.) as a variety of structures are expected to be used in advanced levels of language.

2.1.2.2 Lexical complexity

Although there have been a number of studies investigating the development of L2 grammatical systems, vocabulary acquisition research has attracted a lot of attention in L2 research relatively recently. However, it is now widely accepted that, along with other variables, there is a significant correlation between the vocabulary knowledge and reading comprehension (Lu, Gamson, & Eckert, 2014; Schmitt, Jiang, & Grabe, 2011). The question that has long been asked in vocabulary learning is about the size of the vocabulary knowledge to comprehend the texts. Research

suggests that the readers should be familiar with around 95% (Laufer, 1989) to 98% (Hu & Nation, 2000) of the words in a text to be able to comprehend the text and increase their vocabulary knowledge. Taking the 98% as the threshold for the words needed to be known to comprehend a text, Nation (2006) further argued that around 8000 to 9000 word-family vocabulary should be known by the reader for unassisted comprehension of the texts. Similarly, Schmitt et al. (2011) showed that although higher knowledge of vocabulary led to better comprehension of the texts, there was not a threshold after which the comprehension dramatically increased, rather there was a gradual increase in the comprehension of the text along with the increase in the vocabulary knowledge. All these results indicate the need for higher size and depth of vocabulary acquired by the learner 'over a more prolonged period of time' (Schmitt et al., 2011, p. 39), which can be related to the construct of lexical complexity of the language the learner is exposed to over the period of learning.

Being a multidimensional construct (Kalantari & Gholami, 2017), lexical complexity is mainly measured through three components, namely, lexical diversity, lexical density and lexical sophistication (Bui & Skehan, 2018; Bulté & Housen, 2012, 2014; Johnson, 2017; Skehan, 2009). Lexical diversity refers to the "range of different words in a text" (McCarthy & Carvis, 2010, p. 381) and has been mostly measured via type-token ratio (TTR). Type refers to each individual and unique word while token refers to total number of words in a text. TTR is calculated by dividing the number of different words, types, to the total number of words in a text, tokens (Malvern, Richards, Chipere, & Durán, 2004). TTR can range from 0 to 1, the higher score meaning higher diversity. Since TTR is sensitive to the text length, the diversity of the text decreases as the number of words in the text increases (Malvern et al., 2004; McCarthy & Jarvis, 2010). In order to avoid the confounding variable of

text length, several other indices for lexical diversity have been designed and utilized. Some common transformed measures for lexical diversity such as Root TTR (Guiraud, 1960) and Corrected TTR (Carrol, 1964) have been widely used to eliminate the effect of text length on the lexical diversity scores. However, considering the limitations of these measures, other measures have also been designed. Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005) and vocd-D measure lexical diversity with little or no effect of text length (McCarthy & Jarvis, 2010). MTLD is calculated as "the mean length of word strings in a text that maintain a given TTR value" (op cit., p. 384). Vocd-D calculates lexical diversity through "a computational procedure that fits TTR random samples with ideal TTR curves" (McNamara et al., 2014, p. 67). The final value ranges from 10 to 100, and lexical diversity increases as the score increases (McCarthy & Jarvis, 2010). Although both indices measure lexical diversity, they provide different information regarding lexical diversity (op cit.).

As for lexical sophistication, it refers to the use of more advanced words in language (Kim, Crossley, & Kyle, 2018). The concept of lexical sophistication is usually associated with the use of less frequent words in language, so the word frequency lists have been widely used to identify the range of the word (i.e., the occurrence of the word across the corpus). The software and programs like Nation's Range (Nation & Heatley, 1994) identify what percentage of the words used in a text belongs to the most frequent words (e.g., most frequent 1000 words) based on GSL/AWL lists (General Service List / Academic Word List). Although there is agreement on the use of frequency as an index of lexical sophistication, recent studies have also provided various domains of lexical sophistication including but not limited to the concepts of concreteness (i.e., the more sophisticated a word is, the

more abstract it becomes), age of acquisition (i.e., the words acquired at an earlier age are processed more quickly than the ones learned at an older age), hypernymy (i.e., specificity of the words) and polysemy (i.e., number of meanings of a word) (Crossley & McNamara, 2012; Kim, Crossley, & Kyle, 2018). The computational tools like Coh-Metrix and TAALES (i.e., Tool for the Automatic Analysis of Lexical Sophistication) provide indices and measures for various dimensions of lexical complexity.

Although there is no agreement on the specific dimensions and measures of lexical complexity, the measures (e.g., lexical diversity, lexical sophistication) are expected to increase as the proficiency of the L2 learners increases. On a similar vein, the lexical complexity of the textbooks across grade levels is expected to increase to provide the sufficient and appropriate input to the learners.

2.2 Textbook evaluation research

Textbooks have been a significant, in some cases the main, component of foreign language instruction for many years despite the arguments against the need for and the effectiveness of them in language education (Jordan & Gray, 2019; Thornbury, 2013). Among the merits of the textbooks are that they enable the language programs to have standardized instruction and that they provide a syllabus and structure for the language programs (Hughes, 2020; Richards, n.d.). Considering the significance of the textbooks in language learning programs, the evaluation of the content of the textbooks is rather crucial (Harwood, 2016). The widespread use of the textbooks in the language programs has resulted in a need for research evaluating the materials and the coursebooks for ELT. The majority of the research on textbooks has evaluated both the commercialized textbooks and locally published coursebooks

used in countries in terms of several aspects including but not limited to the content presented, the structure, the types of tasks/exercises, the representation of the culture and the appropriateness of the books to the needs of the learners (Littlejohn, 2011; Tomlinson & Masuhara, 2013). Several checklists have been created to evaluate these and many other aspects of the coursebooks (Byrd, 2001; Tomlinson & Masuhara, 2013). There have also been a number of studies which investigated the complexity of the texts presented in the coursebooks by using both traditional and recent readability formulas for L1 and L2 and measuring the linguistic complexity of the texts for the learners.

2.2.1 Readability research in L2 textbooks

Readability of the texts in textbooks have been commonly predicted through traditional readability formulas. Many publishers consider at least one kind of formulas while designing the books (Dufty et al., 2006). Several studies used readability formulas to measure the text difficulty of texts in both L1 and L2 English textbooks based on grade levels.

Morales (2019) investigated the readability of texts in textbooks provided by the government and used in public and subsidized high schools in Chile. The study examined whether the readability of the texts decreased as the grade level increased. 8 texts were randomly chosen from the textbooks of each level (i.e., 9th, 10th, 11th and 12th) for the analysis. The texts were analyzed using three readability formulas, which were Flesch Reading Ease, Flesch-Kincaid Grade Level, and Coh-Metrix L2 Readability Index, via Coh-Metrix software. The results showed that there was not a gradual increase in the difficulty of the texts from grades 10 to 12. The texts in the 11th grade textbook were significantly easier to read than the texts in grade 10 and

12. There was also no significant difference between the readability of the texts for 10th and 12th grade textbooks. The overall results showed that the readability of the texts for high school did not gradually decrease, which calls for further analysis of the textbooks and research in the field of language learning.

Another study which investigated text complexity through computerized tools for readability was conducted by Bush, Koons and Sanford-Moore (2016). The study investigated the variation of text complexity both within and across grade levels in government-approved textbooks used across grade 5 to grade 12. A corpus of 1506 reading passages was compiled from 31 textbooks. Text difficulty was measured using the Lexile Framework, which measures "an individual's reading ability or a text's complexity (or, difficulty) followed by an "L" (for Lexile)" (p. 2) based on semantic and syntactic factors. The complexity of the textbooks was then analyzed both within the grade by taking the percentiles of text complexity distributions for 25th, median and 75th and across the grades by measuring the overall text complexity of all the passages for each textbook to have a single score for each grade. For both measures, an L (Lexile) score was obtained to compare the complexity of the texts or textbooks. The results of the descriptive statistics showed that text complexity increased within and across textbooks although the greatest increase was seen during the transition from grade 9 to 10. Although there was a gradual increase across levels, the textbooks for grade 9 were easier than expected while the texts for grade 10 were more complex than expected considering the ranges of L scores and increases across levels.

2.2.2 Linguistic complexity in L2 textbooks

The studies investigating linguistic complexity of the texts in textbooks operationalized complexity mostly related to the aspects of vocabulary/lexis and syntactic complexity. Several studies investigated the syntactic complexity of graded textbooks that were used in the schools or language programs across levels. One of the studies investigated grammatical intricacy in the texts and was conducted by To (2017). Adopting a Systemic Functional Linguistics (SFL) approach by Halliday (2008), To (2017) investigated the grammatical intricacy across 4 textbooks of varying levels (i.e., elementary, pre-intermediate, intermediate and upper-intermediate) in a book series used in the higher education in Vietnam. Grammatical intricacy was defined as "how lexical items are scattered in strings of clauses in clause complex" (To, 2017, p. 129). The study focused only on the reading comprehension of the texts; therefore, only the written texts in the textbooks were used for the analysis. Out of 32 reading texts, 24 texts, 6 from each book, were chosen and used for the analysis. Grammatical intricacy was measured by dividing the total number of ranking clauses to total number of clause complexes, and a high result indicated a complex text. The findings showed that the grammatical intricacy of the books increased as the grade levels increased and they matched the levels of the books; however, the results were not statistically significant. Another finding was that the highest-level book (i.e., upper-intermediate) was not the most grammatically intricate textbook as it had the same intricacy score with the pre-intermediate level book. Since the study only focused on the grammatical aspect of text complexity, To (2017) argued that other language features such as lexical density may have an effect on the complexity of the texts as well. Building on the previous research, To (2018) investigated the linguistic complexity of a textbook series which was widely used at

higher education in Vietnam across four levels. 24 written texts from two different genres, which were Explanations and Reports, were chosen and analyzed for the purposes of the study as it focused on the written language. For the data analysis, linguistic complexity was examined through three constructs, namely lexical density, nominalization and grammatical metaphor. Lexical density was measured by dividing the number of lexical items to the total number of raking clauses. Nominalization was measured by the number of verbal and adjectival nominalization. Grammatical metaphor, which was defined as replacing one grammar class or structure by another, was limited to the cases of "ideational metaphor", which "helps construe the ideational meaning" (p. 5). The results of the analysis showed a statistically significant increase in the rate of lexical density and nominalization of the books from elementary to intermediate level; however, there was no statistically significant difference between the intermediate and the upper-intermediate level books in any measures. This finding was in line with results of the previous study by To (2017) in that both lexical and syntactic complexity of the books increased along with the levels of the books.

Putra and Lukmana (2017) investigated text complexity within and among three different books used in three consecutive grades at senior high schools (i.e., grades 10, 11 and 12) in Indonesia. The textbooks were approved by the Ministry of Education. Three texts from initial, middle and last chapters of the books were chosen from each grade, and differences of text complexity in these 9 texts were analyzed both within and across the grade levels. Text complexity was operationalized through lexical density, lexical variation and grammatical intricacy following Halliday's (2008) SFL Approach. Lexical density was measured by the proportion of total lexical words to total number of ranking clauses. Lexical variation

was measured by dividing the number of different lexical words to total number words. Grammatical intricacy was measured by dividing the total number of ranking clauses to total number of clause complexes. The descriptive statistics of the analysis showed that lexical variation and grammatical intricacy increased as the grade levels increased although there was not a consistent increase within the books. Similarly, lexical variety of the textbooks also increased from grade 10 to grade 12; however, there was a slight decrease in the texts of grade 11. Overall, the results showed that there was a gradual increase in the complexity of the texts as the grade levels of the books increased. Following a similar procedure, Mulyanti and Soeharto (2020) investigated the text complexity within and across three EFL textbooks approved by the Ministry of Education to be used in junior high schools (i.e., grades 7, 8 and 9) in Indonesia. Similar to Putra and Lukmana (2017), 6 texts were selected to be analyzed in terms of lexical density, lexical variety and grammatical intricacy. The descriptive statistics showed that lexical density and lexical variety increased from book 7 to book 8 but decreased in book 9, which was the highest level. As for the grammatical intricacy, it increased from grade 8 to 9, which was interpreted as a decrease in text complexity. Overall, the results of the study indicated that the EFL books used in junior high school did not gradually increase in text complexity unlike the books used in senior high school (Putra & Lukmana, 2017).

Other studies investigated linguistic and text complexity via computer-based tools and readability formulas. Building on the previous studies in Taiwan, which mainly investigated the complexity of textbooks either through frequency wordlists or the readability formulas, Chen (2016) analyzed the text difficulty in the book series in terms of both their lexical coverage of frequency-based word lists and the structural complexity through a variety of readability formulas. The study

investigated whether the text complexity of the textbook series increased through the grave levels both in terms of lexical and syntactic complexity. A corpus of 206 reading texts from three textbooks series that were widely used at high schools in Taiwan was compiled for the study. Each series of books targeted the learners from CEFR A1 to B1 as the students were expected to move to B1 when they complete the high school. For the analysis, the lexical complexity was measured through frequency-based word lists; therefore, the corpus was first tagged for the parts-of-speech (POS) using Python Natural Language Toolkit (NLTK), which was similar to the BNC CLAWS5 tagset. The corpus was then analyzed regarding the vocabulary coverage of top thirteen 1000-word lists form BNC word lists. The percentage of the coverage for each book was measured. For grammatical complexity, several readability formulas were used, each of which assigned a grade level to the texts using different parameters for structural complexity. Principal Component Analysis (PCA) was further used to convert the measures into mutually independent principal components. A cluster-based algorithm, variability neighbor clustering, which identifies the "diachronic linguistic development" in the corpora, was also used as a part of the statistical analysis. The results showed that out of three book series, only one textbook series gradually increased in terms of text complexity apart from a slight deviation in one stage in terms of the decrease in lexical complexity. The other books either showed an increase only in one of the measures or showed a decrease in both measures as the levels increased, which implied that as the levels of the books increased, the texts did not necessarily become more complex.

These studies investigated the linguistic or text complexity through readability formulas and specific lexical and syntactic measures, and they investigated whether the complexity increased across the grade levels. A recent study

by Jin, Lu and Ni (2020) investigated the differences among adapted EFL teaching materials used in different grade levels in China regarding their syntactic complexity. A corpus of 3,368 texts of various length from 12 grade level textbooks (i.e., form grade 1 to 12), which were approved by the Chinese Ministry of Education were collected for the data analysis. L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010) was used to analyze the syntactic complexity of the texts. Syntactic complexity was operationalized through several subconstructs including "clausal coordination, clausal subordination, nonfinite elements, phrasal coordination, and noun phrase complexity" and "overall sentence complexity, overall T-unit complexity and elaboration at the clause level" (p. 197). The results showed that there were statistically significant differences among the adapted textbooks for different grades for all 8 measures of syntactic complexity used for 8 different subconstructs, and the texts increased in their complexity as the grade levels increased. This was an indication that the texts were adapted for each grade level considering different dimensions of syntactic complexity. There was also increase and stability in different patterns at lower and higher grades, which was in line with previous research. This shows that the dimensions of syntactic complexity do not necessarily increase in all dimensions linearly, but text and different dimensions of syntactic complexity at different clusters of grade levels should be considered during text adaptation. The results of the study provide important insights for the complexity research in L2 textbooks.

2.3 Educational policies in Turkey

Before presenting the studies evaluating and investigating the effectiveness of the textbooks in Turkey, it is important to introduce the policies regarding the textbook

preparation and distribution process and foreign language education, more specifically L2 English teaching in Turkey.

The textbooks used in state schools have been distributed by the Ministry of Education in Turkey since 2003, and the teachers and students are required to use these textbooks and their additional supplementary materials in the lessons. These textbooks are authorized by the Ministry of Education and can be published by both the Ministry of Education and/or the private publishers. However, all the textbooks must be approved by the Board of Education and Discipline of Ministry of Education (Talim ve Terbiye Kurulu). The drafts of the books that are presented to the board and selected among the others are examined and approved by a committee of 6 people called panelists, who are the experts in the field, linguists and the graphic designers. The experts in a field are supposed to have a teaching experience of 5 years or a doctor's degree in their fields. The textbooks that are eventually approved by the board are listed to be used in the state schools for five years (MoNE, 2012).

As for L2 English education in Turkey, a more communicative approach has been adopted for language teaching for more than two decades (Kirkgöz, 2007), which eventually affected and required a revision of the curriculum and the textbooks used. The overall curriculum for English teaching further follows the principles of the Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR) (MoNE, 2018a, 2018b). The students at the primary and low secondary levels (i.e., 2nd to 8th grade) are expected to improve their communicative competence. There is a major focus on listening and speaking in the early grades (2nd through 4th grade) as the students are exposed to the oral language through songs, games and other activities. There is very limited use of reading and writing, and the lexical input starts with the cognates and builds on them

with very basic level vocabulary. Reading and writing skills are introduced more from 5th to until 7th grade; however, the focus is still limited. After 7th grade, some focus is given to the written skills although the speaking and listening skills still have the primary focus. The rationale for the design of this curriculum is the transition from cognitively low demanding activities to more cognitively demanding ones (MoNE, 2018a) with an increase in the familiarity of the concepts used. The students are expected to have a proficiency level of CEFR A1 after grade 5 and CEFR A2 when they complete grade 8 (see Table 1). They are expected to be able to talk about concrete subjects and express their basic needs with more basic expressions and ask and answer simple questions to an interlocutor when they have A1 level proficiency (MoNE, 2018a). As they progress through A2 level proficiency, they are expected to understand basic expressions fundamental to communication and able to communicate about topics familiar to them (MoNe, 2018a.)

As for the curriculum for secondary education (i.e., high school), the overall curriculum aims to increase the communicative competence of the learners with an emphasis on the integration of four skills and the academic language. Similar to the primary school education, the authenticity and the natural use of the language is emphasized. The proficiency of the students is expected to increase as they move across the grades, starting with A1/A2 level in grade 9 and moving beyond B2+ after completing grade 12 (see Table 1). Although the grade 9 is seen as a revision for the level of A1 and A2, it has been stated that the vocabulary and the structures used in the textbooks can be more advanced than the language presented in the primary and lower secondary education. Therefore, it can be argued that an increase in the complexity of the language is expected as the students move from grade 8 to grade 9.

Similarly, as the grade levels increase, students are expected to be exposed to more advanced linguistic input (see Table 1).

Table 1. English Language Curriculum in Turkey

CEFR Levels / Hours per week	Grades	Skill focus
A1 2 hours	2	Listening & Speaking
	3	Listening & Speaking Very Limited Reading and Writing
	4	Listening & Speaking Limited Reading Very Limited Writing
A1 3 hours	5	Listening & Speaking Limited Reading Limited Writing
	6	Listening & Speaking Limited Reading Limited Writing
A2 4 hours	7	Primary: Listening and Speaking Secondary: Reading and Writing
	8	Primary: Listening and Speaking Secondary: Reading and Writing
A1/A2 4 hours	9	All four skills integrated with an emphasis on Listening and Speaking
A2+/B1 4 hours	10	All four skills integrated with an emphasis on Listening and Speaking Limited focus on Language Structures
B1+/B2 4 hours	11	All four skills integrated with an emphasis on Listening and Speaking Limited focus on Language Structures
B2+ 4 hours	12	All four skills integrated with an emphasis on Listening and Speaking Synthesis of Language Structures

Source: [MoNE, 2018a]

All of these suggest that the linguistic complexity of the language presented to the students in the textbooks prepared, approved and used for L2 English

education in state schools in Turkey is expected to increase as the grade levels and the respective CEFR levels increase. Whether the textbooks meet the criteria for the language to be used, topics, themes and functions of language to be covered and values introduced in the textbooks is another question to be investigated.

2.3.1 L2 Textbook evaluation research in Turkey

Majority of the textbook evaluation research regarding English teaching materials in Turkey has investigated the textbooks in terms of various aspects including pragmatics, more specifically in terms of Grice's maxims of conversation (Arıkan, 2007); sociolinguistic aspects such as the values presented in the textbooks (Aslan, Keskin, & Önder, 2019), frequency of culture-specific elements (Çakır, 2010; Ugurlu & Tas, 2020) and representation of national and global identity (Köroğlu & Elban, 2020). Other studies analyzed the books in terms of their overall effectiveness regarding various aspects including the activities and exercises presented in the books (Tok, 2010; Tüm & Emre-Parmaksız, 2017); perceptions of learners and teachers for the effectiveness of the books (Şener & Mulcar, 2018; Şimşek & Dündar, 2016; Tekir & Arıkan, 2007; Tok, 2010). These studies mainly focused on aspects other than the language, readability and the complexity of the texts in the textbooks. Regarding the readability of the texts in the textbooks, several studies investigated readability across grade levels in L1 Turkish textbooks of various subjects including science (Karakus, Aydın, & Hallac, 2015), Maths (Cetinkaya, Yenmez, Celik, & Ozpinar, 2018) and social sciences (Yılar, 2020).

Several studies examined the readability of L1 Turkish textbooks used in Turkish schools through readability formulas or traditional readability indices (Batur & Ozcan, 2020; Mert, 2013; Turkben, 2019). Mert (2013) investigated the

readability of narrative and informative texts in L1 Turkish textbooks used in middle schools. Cetinkaya-Uzun Readability Formula (2010), which was specially designed for predicting readability of Turkish texts, was used to measure readability. Similar to traditional readability formulas, the formula assesses readability through average sentence length and average word length. Turkben (2019) also investigated the readability of the narrative and informative texts in middle school L1 Turkish textbooks using the readability formula of Atesman (1997), which was adapted from Flesch Reading Ease formula to Turkish, and the Cetinkaya-Uzun (2010) formula, both of which assigned a readability score to the text through measuring average sentence length and average word length.

Although several studies investigated L1 and L2 Turkish textbooks in Turkey, very few studies assessed the readability of L2 English textbooks used in Turkey. As a part of her qualitative study, Kalayci (2018) compared the readability of CEFR B1 and B2 L2 Turkish (i.e., Yeni Hitit Series) and L2 English (i.e., Efekta General English) textbook series available in Turkey. Flesch Reading Ease formula and the adopted version of Flesch formula by Atesman (1997) were used to predict the readability of the English and Turkish textbook series respectively. In addition, the length of the texts was measured through the number of letters. The results showed that the Efekta textbook series included texts with a higher variety of difficulty levels despite the fewer number of texts as opposed to Yeni Hitit Series. This study is of significance in that it is one of the few studies that investigated the readability of English textbooks available in Turkey using traditional readability formulas.

To the best of my knowledge, no studies have investigated the readability of the L2 English textbooks that have been used in Turkish state schools and that were

published and provided by the state through the use of readability formulas or other readability indices. Similarly, few studies have examined the linguistic complexity (e.g., vocabulary coverage, syntactic complexity) of the English textbooks used in K12 education in Turkey.

Regarding lexical coverage of the textbooks, Gedik (2020) compared the use of motion lexicon in high school textbooks used in public schools and the English university entrance exams. Two corpora were used for the study. One corpus consisted of the coursebooks and workbooks currently used in public high schools from grade 9 to 12, which had 301,255 words while the second corpus consisted of English university entrance exams from years 2010 to 2019, which had 66,913 words. Two corpora were compared regarding the frequency of motion verbs, namely manner verbs, and path verbs. The results showed that there was a statistically significant difference between the two corpora regarding the use of manner verbs in that the textbooks included higher number of manner verbs. However, no statistically significant difference was found between two corpora regarding the frequency of path verbs. One limitation of the study was that the results might have resulted from the differences in the number of tokens in each corpus. Another study that investigated the vocabulary coverage of the textbooks was conducted by Asik (2017), who analyzed the language in the locally published coursebooks, in terms of the frequency of the target vocabulary. 21 pre-service teachers studying at a state university in Turkey were trained on how to use corpora and frequency lists to analyze the vocabulary coverage of the textbooks. The students then analyzed the frequency and use of the words in the wordlist of the textbook series used for secondary schools, which corresponded to CEFR A1, A2 and B1 levels. The students analyzed the textbooks both quantitatively (i.e., the frequency of

the words) and qualitatively (i.e., the selection, presentation and practice of the words in the books). The results showed that 70% of the words in the lists were among the most frequent 5000 words in COCA although the frequency of the words decreased as the proficiency level increased. There were also some concerns regarding the selection, presentation and practice of the words in the target lists.

Ağçam and Babanoğlu (2018) investigated the differences between the English textbooks used in public secondary schools at grade 6 in Turkey and Germany. The study compared the complexity of the textbooks in terms of their lexical density and readability. For the study, two sets of corpora consisting of reading comprehension texts were compiled from each textbook, which consisted of 1,221 words and 2,240 words for the textbook used in Turkey and Germany, respectively. The data was then analyzed for lexical density and readability through a computational tool named 'textylizer' (http://textalyser.net/). The results were further analyzed using Fog Scale, which assigns a higher score to short sentences and a lower score to more complex sentences. The results showed that the reading comprehension texts in both textbooks had high lexical density although the textbook used in Germany had less density than the other. Regarding readability, both textbooks were suitable to the levels of the students while the textbook used in Germany was less readable. The textbooks were also analyzed for the increase in text complexity within the book. The results showed that the texts used in Turkish books had a "steadier curve" (p. 956) than the texts in the textbook used in Germany. The study showed that the textbooks had texts with high lexical complexity and readability, which can facilitate the learning process. The study only focused on the progression of text complexity within the books and across two books of the same

level; it did not investigate the complexity of the texts across different grades or levels of English.

When the significance of L2 English textbooks, especially in terms of providing input to the learners in EFL contexts like Turkey, and the limited number of studies investigating the effectiveness of the L2 English textbooks used in Turkey are considered, it can be claimed that there is a need for the evaluation of the textbooks used in the state schools, especially regarding the difficulty and complexity of the language presented to the leaners in order to be able to provide the learners with the input suitable to their levels of proficiency.

CHAPTER 3

METHODOLOGY

This chapter presents the methodology of the present study. It first states the purpose of the study and the research questions designed for the study. Then it describes the corpus of texts compiled and used for the study, the criteria for the selection of the texts and the tools used. It further presents the measures and indices used to analyze the texts. The chapter concludes by presenting the statistical analyses used to investigate the research questions.

3.1 Research questions

Building on the gap in the literature investigating linguistic complexity and the development of linguistic complexity across grade levels in L2 English textbooks used in Turkish state schools, the present study investigates text difficulty and linguistic complexity of L2 English textbooks approved by Ministry of Education and currently being used in Turkish state schools. It further compares the linguistic complexity of two locally published textbooks to two comparable textbooks from a well-known and internationally used series of textbook (i.e., Cambridge Interchange3 and Cambridge Passages1) to investigate whether the texts in these textbooks are similar in terms of their linguistic complexity. The study is motivated by the following research questions:

1. Is there a gradual increase in the text difficulty/complexity of reading texts in the textbooks for different grade levels (grades 2-12)?

   1.a. Is there a gradual increase in the readability scores of the reading texts across grade levels?

1.b. Is there a gradual increase in syntactic complexity of the reading texts across grade levels?

1.c. Is there a gradual increase in the lexical complexity of the reading texts across grade levels?

2. Is there a gradual increase in linguistic complexity (i.e., syntactic complexity and lexical complexity) of the listening/audio texts for different grade levels (grades 2-12)?
3. Is the linguistic complexity of the texts in two locally and two internationally published comparable textbooks similar?

   3.a. Do the two locally and two internationally published comparable textbooks have similar overall readability levels?

   3.b. Do the two locally and two internationally published comparable textbooks have reading texts with similar linguistic complexity levels (i.e., syntactic complexity and lexical complexity)?

   3.c. Do the two locally and two internationally published comparable textbooks have listening texts with similar linguistic complexity levels (i.e., syntactic complexity and lexical complexity)

It is expected that the texts taken from different grade levels and corresponding to different CEFR levels show a gradual decrease in overall readability, which is measured through the readability formulas, and an increase in the lexical and syntactic complexity measures as the grade level and/or the CEFR levels increase. Similarly, the linguistic complexity of the listening texts for different grade and CEFR levels are expected to gradually increase with the increase in the CEFR levels. In addition, it is also assumed that the reading and listening texts in the

locally and internationally published textbooks of the same or approximate CEFR levels are similar in terms of their readability and linguistic complexity.

3.2 Corpus for the study

In order to test the hypotheses of the present study, a corpus of reading and listening texts in one set of textbooks used in the state schools from grade 2 to 12 in Turkey and two textbooks from an internationally published textbook series were compiled.

3.2.1 The textbooks

The data for the study consisted of a specialized corpus of texts compiled from a series of locally published textbooks that are used in the state schools in Turkey starting from grade 2 to grade 12 and two textbooks from an internationally published textbook series (i.e., Cambridge Interchange3 and Passages1) (see Appendix A for a complete list of the textbooks). The textbooks which were approved by Ministry of Education and have been used in state schools were accessed through the online education network of the Ministry of Education (i.e., EBA). These textbooks have been in use approximately for two or three years. For each grade level, one textbook and its accompanying workbook (if present) were used for the present study. The textbooks used for the high school level were selected through the examination of the EBA videos while the textbooks for primary and secondary schools were decided based on the responses to a survey collected from a group of teachers currently teaching at state schools. The Interchange3 (Richards, Hull, & Proctor, 2017) and the paired Passages1 (Richards & Sandy, 2015) series, which were published by Cambridge University Press, were used to compile the corpus for the internationally published and used textbook materials. More

specifically, the Interchange3, which corresponded to CEFR B1/B1+ and Passages1, which corresponded to CEFR B2, were selected to compare with the textbooks used in 11th and 12th grades in Turkish state schools. The internationally published textbook series was chosen because of several reasons. First, the researcher had access to these textbooks. In addition, these textbooks were prepared by Professor Jack J. Richards, who has specialized especially in material design and foreign language education. These textbooks were further developed by considering the feedback from around 1,500 teachers around the world. The target audience for the internationally published textbooks was young-adults and adults, which would match the target audience for the state textbooks.

The books used for the compilation of the corpus included the student's book in the primary and secondary school. The textbooks in these grades are not accompanied by a separate workbook; however, they contain pages for further self-study or practice for the students (except for the textbook used for grade 6), which may be claimed to function as workbooks. Therefore, the self-study/practice parts of the books in the primary and secondary school and the separate workbooks at high school level were included in the corpus as well.

The textbooks for the primary school grades 2-4 mainly focused on the listening and speaking skills, and there was very limited focus on reading and writing (e.g., word and sentence-level written input as a part of short dialogues and exercises) as stated in the curriculum designed by the Ministry of Education (MoNE, 2018a, 2018b). Similarly, there was limited focus on the reading and writing skills in grades 5-6 with relatively short texts, which corresponded to CEFR A1 level. However, the textbook for grade 5 differed from the others in that it mainly presented texts with a focus on both listening and reading in that the students are

expected to read and listen to the texts at the same time. The focus on the reading and writing skills increased with the grade levels 7 and 8 as the proficiency level increased to A2. In the high school textbooks, there was focus on 4 skills, and the texts were longer.

Due to the lack of written texts in the textbooks for the primary levels and the focus on the listening and speaking skills, the texts for listening were also included in the study in order to include the lower-level textbooks (e.g., grades 2, 3 and 4) in the analysis. A corpus of texts with two sub corpora (i.e., reading texts and listening/audio texts) were compiled for the present study. Several criteria were considered for the selection of the texts and the compilation of the corpus, which remained the same for the compilation of texts from each book.

3.2.2 The selection of texts for the written corpus

The written corpus consisted of the reading texts in the textbooks in order to be consistent with the previous studies on readability and grading language materials. The written corpus was used to measure the difficulty of the texts through the readability formulas and the linguistic complexity measures, which affected the criteria for identification and the choice of the texts. For each grade level, the curriculum and the objectives of each unit were analyzed, and these objectives were used to identify the texts for specific skills in the textbooks.

Since there was a very limited focus on reading and writing across grades 2 to 4, only 10 texts were selected from the textbook for grade 4, which is very limited in number. Therefore, the written corpus consisted of reading texts ranging from grade 5 to grade 12. Since grades 5 and 6 correspond to CEFR A1 levels (MoNE, 2018a),

the data could still give information about the readability of lower-level textbooks with low CEFR levels.

In order to identify the reading texts accurately, the objectives of each unit in the curriculum were analyzed for the grades. However, one problem with the objectives in the curriculum of the primary school grades was that the objectives did not always meet the criteria developed by Mager (1975), which emphasized focus on performance, conditions and criterion (Brown, 1995). The criteria explained "what the learner will be able to do", "important conditions under which the performance is expected to occur" and "the quality or the level of performance that will be considered acceptable", respectively (as cited in Brown, 1995, p. 74). In the curriculum designed for English lessons, some of the objectives did not specify a performance or an outcome, but used a language based on comprehension as in the example "Students will be able to understand information about important places." (MoNE, 2018a). Since the objective was too broad and not based on performance or a specific type of text or condition, the texts that meet these objectives were included as reading texts, which was a problem for grade 5 as the textbook did not always specify the specific skill on which a text or activity focused. In addition, especially in the lower grades, some of the texts focused both on listening and reading (e.g., reading passages or dialogues other than speech samples for speaking activities). In such cases, if the purpose and the skill of the activity the text was used for was clear, which was identified through the instructions, objectives and overall purpose of the text, then it was included either in the written or audio corpus, accordingly. Otherwise, it was included in both sub corpora. The number of such texts were not high, though. Similarly, the texts that were used as a prompt for other skills (e.g., speaking) and input for grammatical structures or language functions were not

included in the corpus as the goal of the texts are not always reading comprehension, and the language might not necessarily reflect the level of the book.

The final written corpus included a variety of text types and passages. The texts used in the textbooks for primary school included various texts such as paragraphs, (picture) stories/narratives, emails, lists of instructions and guidelines, invitation cards, news articles, diaries, brochures, blog posts, recipes and letters. One very common type of text in the primary school textbooks was the dialogues, and they were also included in the corpus. However, the names indicating the turns of the speakers were omitted. The texts for the high school included types of texts similar to the primary school textbooks, but there were also a variety of texts like essays, website posts and various other informative, descriptive and narrative texts that differed in length.

The written texts were mainly taken as they are presented in the textbooks, and no changes were made except for the cases when some texts were put in order or required a completion of the activity (e.g., filling the blanks with the sentences taken from the text) for comprehension. However, all the other elements that are not inherently related to the content of the text were excluded, which included the instructions for the activities about the text, the graphs or pictures related to the text or the exercises (e.g., comprehension questions, discussion questions, vocabulary exercises). The texts like surveys and questionnaires, the literary texts including poems, charts and graphs were also excluded. The titles, headlines and captions, names of the authors were further deleted and excluded from the corpus (Leander, 2006) so that the main input was the text itself. The sentence-, phrase- and word-level input, except for the sentences or facts used for reading comprehension in especially lower levels, was also excluded from the study. These included sample

sentences, lists of facts, brochures and ads without sentences, vocabulary lists, single sentences on pictures, dialogues for speaking samples or texts as a part of activities.

3.2.3 The selection of texts for the listening/audio corpus

The audio corpus consisted of texts compiled from the textbooks from grade 2 to 12. The texts were mainly extracted from the supplementary resources for the teachers via EBA (i.e., Teacher's Book, audio transcripts). In cases when the audio transcripts were not present as the textbooks did not include a teacher's book for some grades in primary school and only the recording, the audio recordings were transcribed by the researcher.

The final audio corpus mainly consisted of dialogues, monologues, commercials, videos, documentaries or other informative texts and various other oral input for listening comprehension. Majority of the listening texts, including lists of sentences for the lower levels as well, were included in the corpus. The comprehension questions for the listening texts, sample sentences/input for pronunciation, answers for exercises and the texts with a major focus on reading were excluded from the audio corpus. The input for introducing the grammar units or language functions (e.g., warm-up activities) were also excluded since the focus of the sub corpus was on the listening comprehension. The songs and chants were included in the corpus only if they had complete sentences rather than phrase- or word-level input. The names indicating the turns in the dialogues and fillers (e.g., hmm, well, oh) were deleted from the texts as well (Lu, 2012).

3.2.4 Corpus compilation

In order to compile the corpus, the PDF documents of the locally published textbooks were downloaded from the website of the online education network of the Ministry of Education (i.e., EBA). The PDF documents were converted into .txt document using AntFileConverter (Anthony, 2021). The texts, which suited the above-mentioned criteria, were later extracted from the converted documents and saved in .txt format. In cases when there were problems with the legibility of the converted documents, the texts were extracted from the PDF files separately and saved as .txt documents.

As for the internationally published textbook series, the researcher did not have access to the PDF files of the textbooks, but the Teacher's books; therefore, the written texts were scanned to the computer through Microsoft Lens and converted into .txt files. The corpus was later cleaned for any errors (e.g., spelling). The transcripts of the audio files were extracted from the Teacher's book and saved as .txt files.

As for the coding of the texts, each text was named with information about the grade it belonged to, type of material, its ranking number in the grade, the skill it focused on and the ranking of the text among the texts for that skill. A sample naming of the files is as follows: 12_SB_003_L_001. Further information about the texts were also recorded on a spreadsheet document including the title of the text, the grade level, the publisher, modality/skill, and the type of the text (e.g., dialogue, email, story, informative text, biography).

A total of 765 texts with a total of 145,914 words were compiled from the textbooks across grades 2 to12 and from two internationally published textbooks for the present study. The written corpus consisted of 329 texts with a total of 82,090

words while the audio corpus consisted of 503 texts with a total of 63.824 words. Table 2 summarizes the details of the corpora according to the grade levels.

Table 2. Details of the Texts in the Corpus

Grade Level	Number of Texts in the Corpus		Total Number of Words		Number of words / Mean		Number of words / SD	
	Written	Audio	Written	Audio	Written	Audio	Written	Audio
2	-	53	-	1,345	-	25.38	-	10.71
3	-	51	-	2,090	-	40.98	-	22.92
4	10	59	270	2,313	27	39.20	12.06	19.02
5	18	29	1,578	1,910	87.67	65.86	33.82	39.37
6	17	29	1,004	3,705	59.06	127.76	27.46	40.07
7	39	29	4,751	2,586	121.82	89.17	43.89	41.50
8	34	35	5,807	4,405	170.79	125.86	62.86	54.01
9	47	21	10,632	4,444	226.21	211.62	74.37	65.51
10	31	18	12,697	4,297	409.58	238.72	137.63	122.02
11	47	21	15,632	7,620	332.60	362.86	95.66	82.73
12	31	18	10,161	3,453	327.77	191.83	128.51	106.88
Interchange3	32	41	11,295	10,806	352.97	263.56	54.019	115.68
Passages1	23	32	8,263	14,850	359.26	464.06	54.909	119.78
Total	329	436	82,090	63,824	249.51	146.39	140.946	146.383

Note. SD = standard deviation

3.3 Measures and indices

In order to investigate the research questions, several measures and indices that measured overall readability of the reading texts and linguistic complexity of the reading and listening texts were used.

3.3.1 Readability formulas

Both traditional readability formulas (i.e., Flesch Reading Ease and Flesch-Kincaid Grade Level) and an automated readability formula (i.e., Coh-Metrix L2 Readability Index) were used to measure the readability of the written texts. Although the traditional readability formulas have been widely criticized for providing a partial look into the readability of the texts, the validity of these formulas have also been shown in especially L1 English research. They have also been widely used in the analysis of graded textbooks. Therefore, these formulas were included in the present study as well, and they were measured through Coh-Metrix. Considering the limitations of and the criticism against traditional readability formulas (e.g., their focus on the measures of average sentence and word length, the accuracy of the results for L2 texts), Coh-Metrix L2 Readability Index, which analyzes the texts both in terms of the sentence and word-level features and the cohesion between and among the sentences (Graesser et al., 2014), was further used for measuring the readability of the texts. Studies investigating the effectiveness of Coh-Metrix L2 Readability Index showed that it performed better at differentiating the readability of texts than the traditional readability formulas (Crossley et al., 2008, 2011).

A minimum number of 200 words were required for an accurate score for the traditional readability formulas on Coh-Metrix while the tool itself required a minimum of 100 words for a text to have a more accurate analysis. These were also considered while analyzing the readability of the texts. Majority of the texts in the primary and lower secondary levels did not have more than 200 words as there were only 9 texts over 200 words in the textbook for grade 8; therefore, only the texts from the textbooks for high school were included in the analysis of readability formulas.

The readability formulas provide a general measure of readability for the texts. However, in order to see how the linguistic complexity of the texts differ across different grade levels, several indices and measures which provide analysis on various levels of language for both syntactic and lexical complexity were used to analyze both written and oral data in the present study.

3.3.2 Syntactic complexity measures

Following Ortega (2003) and Bulté & Housen's (2012) multidimensional approach to L2 complexity, several measures that analyzed the texts in various aspects were used to measure syntactic complexity of the texts. Therefore, a total 11 measures were employed from both tools (i.e., Coh-Metrix and L2SCA) to measure syntactic complexity.

6 syntactic complexity measures provided by Coh-Metrix (Graesser et al., 2014) to measure syntactic complexity were used in the present study. In addition, the descriptive index of "mean length of sentences" (READSL) was also included in the analysis as it is one of the most common measures used for syntactic complexity. The first measure used was the measure of "the mean number of words before the main verb, or left embeddedness (SYNLE)" (Graesser et al., 2014, p. 70), which increases as the number of words before the main verb increases. The sentences with higher amount of left embeddedness are claimed to become more syntactically complex as they contain more higher-level constituents such as embedded clauses (Graesser et al., 2014; Graesser et al., 2004). In addition, as the "average number of modifiers per noun phrases (SYNNP)" (Graeasser et al., 2014) increases in a text, the text becomes denser in noun phrase (NP), which affects the syntactic complexity of the text and the increases the cognitive load. Therefore, SYNNP was another

measure used in the study. Two other indices measured the minimal edit distance between two consecutive sentences in terms of the parts of speech (POS) and words, which are coded as SYNMEDpos and SYNMEDwrd on Coh-Metrix, respectively. These indices measure how much each consecutive sentence should be edited to have the same syntactic construction or the same meaning. The indices show how "dissimilar" two sentences are, and especially the measure of SYNMEDpos was found to be correlated with syntactic complexity (Graesser et al., 2014). On a similar vein, Coh-Metrix measures the similarity of the sentence structure both between the adjacent sentences (SYNSTRUTa) and across all combinations of sentences (SYNTRUTt) in a text. Both measures were used in the present study as the texts are expected to be more difficult to process as the similarity of the sentences in a text decreases (Crossley et al., 2008; Graesser et al., 2014).

In addition to the measures provided by Coh-Metrix, several traditional measures, which have been widely used to measure syntactic complexity in L2 research, were used in the analysis of the texts. A variety of measures tapping into different subdimensions of syntactic complexity were selected and used based on previous research on complexity (Bulté & Housen, 2012; Jin et al., 2020; Ortega, 2003) and measured via online web-based L2 Syntactic Complexity Analyzer (L2SCA, Lu 2010). The dimensions included overall syntactic complexity, subordination and subclausal complexity as suggested by Norris and Ortega (2009) and coordination and noun phrase complexity (Jin et al., 2020; Yang, Lu, & Weigle, 2015).

Overall syntactic complexity was measured through mean length of sentence (MLS) and mean length of T-unit (MLT). Although these two measures may be affected by various factors, they still prove to be significant indicators of global

syntactic complexity (Lu, 2011; Jin et al., 2020). MLT is one of the most common measures used for measuring overall syntactic complexity in L2 research (Norris & Ortega, 2009; Johnson, 2017). In addition, MLS and MLT were both found to be a significant indicator of the increase in the syntactic complexity in the textbooks used across different grades in China (Jin et al., 2020). The study by Jin et al. (2020) provided further important insights into the use and selection of syntactic complexity measures for the present study. Following Yang et al.'s (2015) framework, Jin et al. (2020) operationalized syntactic complexity on different subdimensions which were measured by various measures. They measured the dimensions of overall sentence complexity via MLS, overall T-unit complexity via MLT, clausal coordination via T-units per sentence (T/S), clausal subordination via dependent clauses per T-unit (DC/T), elaboration at clause level via mean length of clause (MLC), phrasal coordination via coordinate phrases per clause (CP/C), noun phrase complexity via complex noun phrases per clause (CN/C) and nonfinite elements/subordination via nonfinite elements per clause (NFE/C) (op cit.). The results showed that all measures showed statistically significant differences across grade levels, MLS and MLT having the largest effect size. However, the results for the analysis for the predictive powers of the measures suggested that 5 measures were the greatest predictors of syntactic complexity, which included both global and fine-grained measures. Although MLT was not among the suggested 5 measures, it was included in the present study along with MLS to measure global complexity because of its large effect size and widespread use in L2 research. The remaining 4 measures included DC/T, T/S, CN/C and NFE/C. Following the framework and the results of the study, 3 of these four measures were also used in the present study. As in Jin et al. (2020), DC/T was selected as the measure for subordination, T/S was used for

operationalizing sentence coordination, and CN/C was used for noun phrase complexity in the present study as well. In addition, MLC was used in the present study to measure "subclausal complexity via phrasal elaboration" as suggested by Norris and Ortega (2009) because it differed from other measure based on length.

Overall, 11 measures of syntactic complexity were included in the present study to analyze different aspects of the construct. Some of the measures (e.g., MLS and READSL) were identical in both tools; however, they were included in the analysis and measured by both tools as there were texts which had fewer than 100 words in the corpus, and these texts could not be analyzed by Coh-Metrix. Therefore, L2 Syntactic Complexity Analyzer was alone used to measure syntactic complexity of these short texts.

3.3.3 Lexical complexity measures

Although there has been a focus on syntactic complexity in L2 research, the significance of lexical complexity measures has recently been recognized. Several studies investigating the proficiency levels of the learners and the readability of the texts across proficiency levels have shown that lexical complexity of the texts play a significant role in identification of different proficiency levels (Sung, Dyson, Chen, Lin, & Chang, 2015; Vajjala & Meurers, 2012). Vajjala and Meurers (2012) found that when lexical and syntactic complexity measures are combined, they classified the texts in a corpus of newspaper articles graded for different levels with 82.3% accuracy while the sole use of lexical or syntactic complexity features would result in a lower accuracy rate (71%). Similarly, in their study of CEFR-levelled texts of Chinese as a Foreign Language, Sung et al. (2015) found that lexical features like "average vocabulary levels" (p. 383) were among the most significant factors for

foreign language learning, and they may affect the difficulty of a text. Considering the results of the studies and the significance of lexical complexity in language acquisition and grading of the texts, various lexical complexity measures were included in the study.

Following the multidimensional approach to lexical complexity, several measures for different dimensions of lexical complexity from both Coh-Metrix (Graesser et al., 2004) and Lexical Complexity Analyzer (LCA, Lu, 2012) have been utilized in the present study. Type-token ratio (TTR) is one of the most common measures used to measure lexical complexity, more specifically lexical diversity. However, since TTR is sensitive to text length (Malvern et al., 2004; McCarthy & Jarvis, 2010), automated measures, which eliminate or alleviate the confound of text length, have also been used. Other most common measures for lexical diversity, VOCd and MTLD (Measure of Textual Lexical Diversity), which are the transformations of type-token ratio, have been proved to be valid measures for lexical diversity as well (McCarthy & Jarvis, 2010). Recent research showed that although both measures aimed to eliminate the effect of text length on the results, among several other measures, MTLD was affected the least by the text length especially when used with shorter texts (Koizumi & In'nami, 2012; Zenker & Kyle, 2021). Considering that there were texts that had around 100 words, and those that ranged between 100 to 200 words, MTLD measure was used to analyze the lexical diversity of the texts through Coh-Metrix. In addition, one lexical diversity measure from Lexical Complexity Analyzer were also added in order to be able analyze the lexical diversity of the texts with less than 100 words. Therefore, a transformed measure of TTR, CTTR (Corrected TTR) was further used as a measure of lexical diversity in order to control the effect of text length on the results to a degree. To

measure the diversity of different categories of words, a lexical word variation diversity measure (i.e., LV) and a verb diversity measures (i.e., SVV1) were calculated through Lexical Complexity Analyzer. The measures of CTTR and SVV1 were among the 9 measures that were suggested by Lu (2012) in that they performed well in differentiating the levels of oral narratives.

Lexical density and lexical sophistication were measured through lexical density (i.e., LD) and lexical sophistication (i.e., LS-II) measures of Lexical Complexity Analyzer. The measure of LS-II was found to produce valid and reliable results for L2 writings (Laufer, 1994; Lafuer & Nation, 1995). As lexical sophistication is also influenced by word frequency (Crossley et al., 2012), word frequency was further measured through two CELEX word frequency log indices by Coh-Metrix. The indices that provide a logarithm of word frequency for all words (WRDFRQa) and average minimum log frequency for content words (WRDFRQmc). A lower score for the indices indicates a high occurrence of low-frequency words in the text. Overall, a total of 8 lexical complexity measures and indices were used for the present study.

The measures and indices measuring the similar dimensions of lexical complexity in either tool were included in the analysis to be able calculate the lexical complexity of the short texts and the listening texts. However, the texts containing fewer than 50 words could not be analyzed in terms of lexical complexity as the tool Lexical Complexity Analyzer requires the texts to contain a minimum of 50 words. The word limit for a more accurate analysis of the texts on Coh-Metrix was 100, which was also considered in the analysis of the texts.

3.4 Analysis of the texts

Each text was analyzed by the researcher using the online versions of the computational tools utilized in the present study. The scores for each measure were saved to a spreadsheet document for each text with further information about the texts including the text title (e.g. 6_SB_001_L_001), grade, the publisher, the type of skill the text focused on (e.g., reading or listening), the type of the text (although the study did not focus on the genres, the information about the types of the texts was included as it might influence the linguistic complexity of the texts), number of words and the specific measures.

As the affordances and the requirements of the tools varied, different sets of data were saved in different spreadsheets to be analyzed separately. Coh-Metrix required a text of a minimum of 100 words for a meaningful and reliable analysis of the text for its indices and measures while a minimum of 200 words were required for the traditional readability formulas (e.g., Flesch Reading Ease). Similarly, a minimum of 50 words were required for LCA to work properly. The texts that did not meet these criteria were removed from the data set for the specific tool. There were no word limits for the L2SCA, so all the texts in the corpus were analyzed by the tool while the number of texts differed across other data sets. A separate set of data was saved for each tool and type of texts (i.e., reading versus listening) and textbooks (i.e., locally versus internationally published) examined. The texts for readability formulas were also analyzed separately.

3.5 Statistical analysis

SPSS Statistics 26 was used to analyze the data. Prior to the statistical analyses, the data was checked in terms of its distribution. Since the data consisted of an

independent variable (i.e., grade levels) with several groups, each group was checked in terms of its normal distribution for each measure employed. Several different strategies have been developed to check the assumptions of normality in a data. The Kolmogorov-Smirnov test and Shapiro-Wilk tests are widely used to check for the assumptions of normality. However, since these tests have certain limitations (Field, 2009) and they do not have enough power (Larson-Hall, 2015), alternative methods like checking the data visually through histograms, QQ-plots and boxplots are recommended. However, since these techniques can be quite subjective, sole reliance on these methods may not provide accurate and conclusive results for normality (Larson-Hall, 2016). Therefore, the assumptions of normality were checked through the skewness and kurtosis values accompanied with the graphs. Following Tabachnick and Fidell (2013), the values for skewness and kurtosis were set within +/-1.5. The data that did not meet the criteria were normalized by winsorizing the outliers in that the outliers were changed to one unit above or below the next highest score (Tabachnick & Fidell, 2013; Field, 2009). The outliers that could not be normalized were removed from the data. Winsorizing and removal of the outliers normalized majority of the data; however, some datasets were not normalized via winsorizing or transformation. Since it is recommended not to winsorize more than 5% of the data (Tabachnick & Fidell, 2013), and the sample size was small for certain grades, nonparametric tests were used to be able to analyze the sets of data that did not meet the assumption of normality.

The datasets that were normally distributed were analyzed through a series of one-way ANOVAs. The effect size of the significant results for one-way ANOVAs was calculated through partial eta square (η_p^2), which was obtained through SPSS. Following Huck's (2012) criteria, an effect size of .01 was considered small, .06

medium and .14 as large. As for the post hoc tests for pairwise comparisons, Gabriel test was used for the data that met the Levene's homogeneity of variances assumption as the sample sizes were slightly different in all comparisons, and Gabriel test had more power in case of unequal sample sizes (Field, 2009). In the cases when the homogeneity of variances was not sustained in the ANOVA results, the results for Welch test was used to check if there were significant differences among the groups, which was followed by the Games-Howell post-hoc test (Field, 2009). These post hoc tests were used because they had control over Type I error (Field, 2009; Larson-Hall, 2016).

Since the sample size was small for many grades, the *p* value was set to .05 to have more power. However, considering the recent criticism towards the sole reliance on the "dichotomous" *p* (Larson-Hall & Plonksy, 2015), the effect size was calculated for the post hoc tests to see the magnitude of the significance between the means. For parametric tests, Cohen's *d* was used to measure effect size (Larson-Hall, 2016), one reason for which was that *r* can be affected by the difference between the sample sizes (Field, 2009). The formula for Cohen's *d* used pooled *SD*s as in the following formula:

$$d = (M_1\text{-}M_2) / \text{Pooled SD}$$

An effect size of $d = .40$ was considered small, $d = .70$ as medium and $d = 1.00$ as large (Plonsky & Oswald, 2014).

As for the datasets that were not normally distributed, the nonparametric equivalence of ANOVA, which is Kruskal-Wallis test, was used. The pairwise comparisons for the significant results of Kruskal-Wallis test were conducted using the Mann-Whitney test. Since it does not control Type I error, Bonferroni correction was applied for the multiple analyses. In order to measure the effect size for the

significant post test results, *r* was calculated by dividing the *z* score of the Mann-Whitney test to the square root of the total number of observations (Field, 2009). Based on the analysis of Plonsky and Oswald (2014), *r* =.25 was considered to be small, *r* = .40 medium and *r*= .60 large. In cases where too many analyses were required, which would inflate the Type I error rate, a Spearman's correlation test was conducted to see if there was a strong positive correlation between the independent and dependent variables.

CHAPTER 4

RESULTS

This chapter summarizes the results of the statistical analyses of the texts in the corpora using different measures for various constructs. It first presents the results for the readability and linguistic complexity of the reading texts and the linguistic complexity of the listening texts in the textbooks used in state schools. Finally, the results of the comparisons between the textbooks used in grades 11 and 12 to two internationally published comparable textbooks are presented.

4.1 Results for research question 1

The first research question investigated the differences between the textbooks used in different grade levels in terms their text difficulty, which was measured by the overall readability of the texts through readability formulas and the amount of linguistic complexity the texts have, which was measured through the syntactic and lexical complexity measures.

4.1.1 Readability of the reading texts

The readability of the texts was measured through both traditional readability formulas and Coh-Metrix L2 Readability Index. The traditional readability formulas require a minimum of 200 words in a text to provide reliable results. Therefore, the texts that met the criteria were included in the analysis to investigate the differences between the grade levels regarding the traditional readability formulas and Coh-Metrix L2 Readability formula. Due to the limited number of texts with a minimum of 200 words in the lower grades, only the textbooks used in high school (i.e., grades

9-12) were included in the analysis. A series of one-way ANOVAs and following post hoc tests were conducted to investigate the differences among the grade levels.

Since the homogeneity of variances was not sustained for the traditional readability formulas, Welch *F* is used to report the results while the *F* for ANOVA is used for reporting Coh-Metrix L2 Readability Index as the homogeneity of variances was sustained. Considering the results of the homogeneity of variances assumption, the Games-Howell post hoc test was used for traditional readability formulas while Gabriel post hoc test was used for Coh-Metrix L2 Readability Index. The results of the one-way ANOVAs showed that there were statistically significant differences between the texts used in different grade levels in high school with large effect sizes; Flesch Ease, $F(3, 62.425) = 12.981$, $p < .001$, $\eta_p^2 = .235$; Flesch Grade, $F(3, 62.357) = 17.481$, $p < .001$, $\eta_p^2 = .281$; L2 Readability Index, $F(3, 125) = 14.062$, $p < .001$, $\eta_p^2 = .252$.

The Games-Howell post hoc test for the traditional readability formulas showed similar results for both formulas. The results of the Flesch Ease Formula indicated that the texts for grade 9 were more readable than grade 10 and grade 12 ($p < .001$) (see Table 3 for descriptive results) with a large effect size for each grade as Cohen's *d* was 1.22 for grade 10 and 1.36 for grade 12 (see Table B1 in Appendix B for a complete list of effect sizes for pairwise comparisons). Although the mean score for grade 9 was higher than grade 11, which indicated higher readability for grade 9, the results were not statistically significant. The results for grade 11 further showed that the texts in the textbook were more readable than the texts in grade 10 ($p = .027$). Although the result was significant ($p < .05$), the effect size was close to medium (Cohen's $d = .68$). A significant difference was also found between grade 11 and grade 12 ($p = .004$, Cohen's $d = .95$). There was no statistically significant

difference between the mean scores of texts for grades 10 and 12 although the texts for grade 10 had a higher mean.

Table 3. Descriptive Results of Readability Formulas for the Texts in Locally Published Textbooks

Grade Level		Flesch Ease		Flesch Grade		L2 Readability Index	
	N	M	SD	M	SD	M	SD
Grade 9	28	76.46	6.10	5.28	1.26	23.56	4.18
Grade 10	31	65.81	10.68	7.73	2.07	18.32	3.79
Grade 11	44	72.61	9.00	6.52	1.71	19.12	4.28
Grade 12	26	60.84	14.92	8.84	2.87	16.79	3.92

Note. M = Mean; *SD* = Standard deviation

The pairwise comparisons between the grades in terms of Flesch-Kincaid Grade Level formula aligned with the results of the Flesch Ease except that grade 9 had a significantly lower mean (see Table 3 for descriptive results) than all the other levels ($p < .05$) with a medium to large effect size ranging from 1.42 for grade 10, .82 for grade 11 and 1.60 for grade 12. The readability of the texts for grade 10 was lower than grade 11 ($p = .048$), but the effect size was close to medium (Cohen's $d = .63$). The texts for grade 11 were more readable than the texts for grade 12 ($p = .003$, Cohen's $d = .98$). However, no significant differences were found between the textbooks for grade 10 and 12.

The results of the post hoc test for Coh-Metrix L2 Readability Index partially supported the results of the Flesch-Kincaid Grade Level Formula as the texts for grade 9 were significantly more readable than the texts for all the grades above ($p < .001$) with large effects sizes as Cohen's d was 1.31 for grade 10, 1.04 for grade 11 and 1.66 for grade 12. However, no significant differences were found among grade levels 10, 11 and 12.

Overall, the results of the traditional readability formulas (i.e., Flesch Ease Formula and Flesch-Kincaid Grade Level) and Coh-Metrix L2 Readability Index showed that the textbook for grade 9 had texts with higher readability levels than the grades above except for the nonsignificant difference with grade 11 for the Flesch Ease formula. While grade 10 and 11 had significant differences for two traditional readability formulas, the close-to-medium effect size requires further analysis. On the other hand, the results of all the formulas showed that grades 10 and 12 were not significantly different from each other in terms of the level of readability the texts had.

Although the results of the analyses for the readability formulas provided some information about the complexity and probable difficulty of the texts across grades, considering the medium effects sizes and the mixed results of some pairwise comparisons, a more in-depth look at the differences between the grade levels in terms of their linguistic complexity would provide a better understanding of the readability or the text complexity of the texts in these textbooks.

4.1.2 Syntactic complexity of the reading texts

Syntactic complexity of the texts was analyzed through two computational tools, namely Coh-Metrix and L2 Syntactic Complexity Analyzer (i.e., L2SCA). The texts with fewer than 100 words were analyzed using L2SCA, while Coh-Metrix required a minimum of 100 words. Since there were very few texts meeting the criteria of 100 words in the textbooks for grade 5 and 6, these grades are included only in the analysis for the measures of L2SCA.

Out of 6 syntactic complexity indices measured by L2SCA, 5 measures had normal distribution while DC/T violated the assumptions of normality. Therefore,

non-parametric tests, namely Kruskal-Wallis Test and Mann-Whitney test, were used to compare the groups in terms of the amount of subordination the texts had. The data for the other measures were analyzed by one-way ANOVA and Games-Howell post hoc test as the homogeneity of variances was not assumed for any measure. Because the homogeneity of variances was not sustained, the Welch *F* is reported in the following results. Both the parametric tests and the nonparametric equivalent indicated that there were significant differences between the grade levels in terms of each syntactic complexity measure with medium to large effect sizes; however, the post hoc tests showed mixed results for some of the measures.

In terms of length-based measures of syntactic complexity, one-way ANOVAs showed significant differences between grade levels with large effect sizes, MLS, $F(7, 92.734) = 49.804$, $p < .001$, $\eta_p^2 = .530$, MLT, $F(7, 91.941) = 49.100$, $p < .001$, $\eta_p^2 = .526$. The post hoc tests for both length measures showed similar results (see Table 4 for descriptive results). Grade 5 did not have a significantly different mean than grade 6 and 7 for either MLS or MLT measures while it had shorter sentences and T-units than all grades above grade 7 ($p < .001$) and the effect size for each difference was large (see Table B2 in Appendix B for a complete list of effect sizes for pairwise comparisons). Grade 6 had a lower mean score than grade 7 for both measures ($p < .05$, Cohen's $d = .97$ for MLS, .98 for MLT) and all the grades above 7 ($p < .001$) with large effect sizes. There was a significant difference between grade 7 and 8 for MLS, and the effect size was medium (Cohen's $d = .78$). However, no significant differences were found between grade 7 and 9, and grade 8 and 9 for MLS. The texts for the grades 7, 8 and 9 were not different in terms of MLT scores, though, while they had significantly lower

scores than all the grades above 9 ($p < .001$) with large effect sizes. No significant differences were found between grades 10, 11 and 12.

Table 4. Descriptive Results of Length-based Measures of Syntactic Complexity for Reading Texts

		MLS		MLT		MLC	
Grade	N	M	SD	M	SD	M	SD
5	18	7.39	1.98	6.61	1.79	6.59	1.84
6	17	6.41	2.02	6.19	1.76	6.19	1.69
7	39	8.82	2.85	8.23	2.35	7.49	1.86
8	34	11.11	3.02	10.20	2.94	9.46	2.65
9	47	10.51	2.37	9.56	1.83	8.03	1.37
10	31	15.05	3.18	13.30	2.65	10.11	1.48
11	47	14.43	2.28	12.94	1.95	8.98	1.31
12	31	16.11	4.76	14.66	4.30	9.63	2.54

Note. MLS = Mean length of sentence; MLT = Mean length of T-unit; MLC = Mean length of clause; *M* = Mean; *SD* = Standard deviation

Although the measure of MLC is considered as a clause measure in the present study, the results related to it are presented with the values of other length-based measures (see Table 4). The groups had significant differences for MLC, $F(7, 89.724) = 16.539$, $p < .001$, $\eta_p^2 = .291$. Although the results showed some similarities with overall length-based measures, there were some differences, one of which is that there was no significant difference between the texts for grade 5 and grade 9 although grade 5 had a significantly lower mean score than all the texts for other grades above 7 ($p < .001$) with large effect sizes. In addition, there were no significant differences between the comparisons of the books for grade 5, 6 and 7. The texts for grade 6 had significantly lower mean for all the grades above 7 ($p < .05$) with large effect sizes. While the texts for grades 7 and 8 were significantly different ($p = .14$), the effect size was medium (Cohen's $d = .85$). The results showed no differences between grades 7 and 9 similar to the results of MLS and MLT. Grade 7 also had significantly shorter clauses than the texts for 10th, 11th, and 12th grades (p

< .05) with close-to-large and large effect sizes. As for the grades above 7, no significant differences were found between grade 8 and the grades above it. Grade 9 had significantly shorter clauses than grade 10 ($p < .001$, Cohen's $d = 1.45$) and grade 11 ($p = .021$, Cohen's $d = .70$); however, no significant differences were found between grade 9 and 12. The texts for grade 10 had longer clauses ($p = .022$) than the texts for grade 11 with a medium effect size (Cohen's $d = .80$). However, the texts for grade 12 were not different from the texts for grade 10 or 11.

As it can be seen in Figure 2, although there were differences in the significance of the results, the distribution of mean scores for length-based measures showed some overlap.

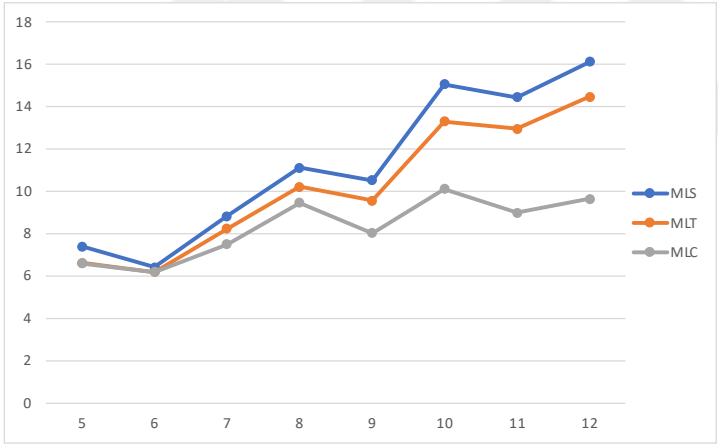

Figure 2. Means of length-based measures across grade levels

As for subordination measure (i.e., DC/T), since the data was not normally distributed, Kruskal-Wallis test was conducted, which showed significant differences among the grade levels, $H(7) = 147.185$, $p < .001$. The Mann-Whitney post hoc test was conducted to measure the differences between the groups. In order to control the Type 1 error, a Bonferroni adjustment was applied, and p was set at .0018. The results mainly showed that the adjacent grades or a cluster of grades did not have significant differences (see Table 5 for descriptive results), providing some support

to the results of the overall length-based measures (i.e., MLS and MLT). The results for the following clusters were not significant: 5-6, 7-8, 7-9, 8-9, 9-10, 10-11, 10-12, 11-12. For the remaining significant results between the grades, *r* ranged from medium (.46) to large (.83). As it can be seen in Figure 3, there was an upward trend for the measure despite the nonsignificant results for some clusters of grades.

Table 5. Descriptive Results of the Remaining L2SCA Measures for the Reading Texts

Grade Level	DC/T			T/S		CN/C		
	N	M	SD	M	SD	N	M	SD
5	18	.02	.03	1.14	.21	17	.31	.19
6	17	.00	.00	1.02	.09	17	.37	.23
7	39	.12	.12	1.05	.10	39	.62	.34
8	34	.14	.15	1.11	.12	34	.92	.51
9	47	.19	.13	1.09	.09	47	.77	.31
10	31	.31	.21	1.13	.10	31	1.11	.34
11	47	.45	.17	1.11	.06	47	.92	.25
12	31	.49	.30	1.09	.08	31	1.03	.47

Note. DC/T = Dependent clauses per T-unit; T/S = T-unit per sentence; CN/C = Complex nominal per clause; *M* = Mean; *SD* = Standard deviation

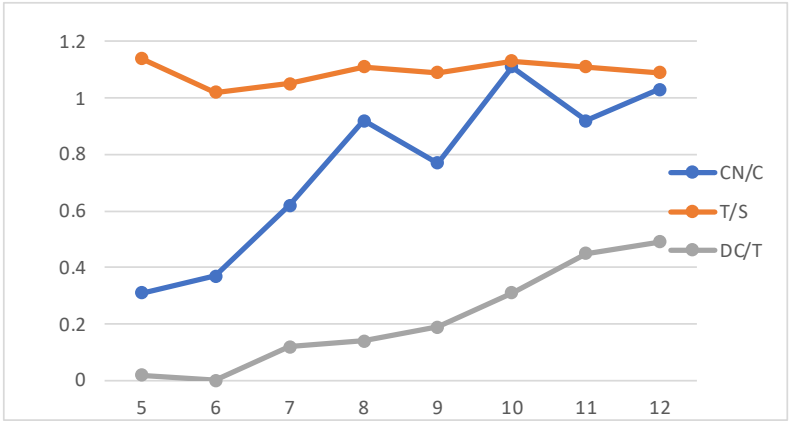

Figure 3. Means of the remaining L2SCA measures across grade levels

Although the result of the one-way ANOVA showed significant differences between grades for T/S ($F(7, 89.163) = 3.551$, $p = .003$, $\eta_p^2 = .078$), few significant

differences were found in the post hoc tests. One significant finding was that grade 6 was significantly different from grade 10 (p = .022, Cohen's d = 1.05) and grade 11 (p = .034; Cohen's d = 1.06). In addition, grade 7 had significantly lower mean scores grade 10 (p = .044) and grade 11 (p = .040); however, the effect size was medium for grade 10 (Cohen's d = .72) and close to medium for grade 11 (Cohen's d = .65). No other significant differences were found between the grades including the lowest (i.e., grade 5) and the highest (i.e., grade 12) grades. As for the results of the measure of noun phrase complexity (i.e., CN/C), the result of the one-way ANOVA showed significant differences across grade levels: $F(7, 92.807) = 27.948$, $p < .001$, $\eta_p^2 = .304$. The pairwise comparisons further showed that although there was no significant difference between grade 5 and 6, grade 5 had a significantly lower mean score than grade 7 (p = .002) and all the grades above ($p < .001$) with large effect sizes. The texts for grade 6 had significantly lower amount noun phrase complexity than both grade 7 ($p < .05$) and all the grades above ($p < .001$) with large effect sizes except for the medium effect size for grade 7 (Cohen's d = .87). There were no significant differences between grade 7 and grades 8 and 9 while grade 7 had a significantly lower mean score than grades 10, 11 and 12 with close-to-large and large effect sizes. As for the pairwise comparisons of the grade levels above 7, the only significant difference was found between the grades 9 and 10 with a large effect size (Cohen's d = 1.03). No other significant differences were found.

The results of the analysis of the syntactic complexity measures of L2SCA indicated that grade 5 and 6 had significantly lower means than all the grades above for the majority of the measures although the two grades were not significantly different from each other. Similarly, the texts for grades 7, 8 and 9 tended to have texts with similar syntactic complexity levels while grades 10, 11 and 12 mostly had

nonsignificant results among themselves. However, there were exceptions to these especially regarding the grade levels above 6 in that some nonsignificant differences were found for some clause-based syntactic complexity measures (e.g., MLC, DC/T, CN/C) between middle school and high school textbooks of different CEFR levels.

The syntactic complexity of the texts was further analyzed by Coh-Metrix measures. Due to the limitations on the word count by the tool, textbooks starting from grade 7 were included in the analysis of the measures of Coh-Metrix, which further resulted into fewer texts for grade 7 ($N = 27$) and grade 8 ($N=30$) while only one text was removed from the analysis for grade 9. All the reading texts for grades 10, 11 and 12 were above the word limit.

A series of one-way ANOVAs were run to compare the mean scores for different groups, and since all the samples violated the assumption of homogeneity of variances, Welch's F is used to report the results. On a similar vein, the Games-Howell post hoc test was used to compare the differences across grade levels.

Out of 7 measures, 5 syntactic complexity measures of Coh-Metrix showed significant differences across grades while there were no significant differences between the grades regarding the measures of SYNMEDpos and SYNMEDwrd. 2 measures, namely READSL (i.e., sentence length) and SYNNP (i.e., modifiers per noun phrase) had somewhat similar results with two measures of L2SCA, which are MLS (i.e., mean length of sentence) and CN/C (i.e., complex nominals per clause), which was expected as these measures have similar underlying concepts. Regarding the results for the measures of READSL and the results of MLS, the only slight difference was that grade 7 ($M = 8.51$, $SD = 2.43$) had a significantly lower mean score ($p < .05$) than both grade 8 ($M = 11.02$, $SD = 3.14$) and grade 9 ($M = 10.31$; $SD = 2.37$). However, the effect size was medium (Cohen's d was .89 for grade 8 and

.74 for grade 9). The remaining pairwise comparisons showed similarity in that no significant differences were observed between grade 8 and 9 while they had significantly shorter sentences than all the grades above ($p < .05$) with close-to-large and large effect sizes ranging from .93 to 1.53 (see Table B3 in Appendix B for a complete list of effect sizes for pairwise comparisons). Similarly, no differences were found among grades 10, 11 and 12. The results for the measure SYNNP showed significant differences between the grades ($F(5, 91.261) = 11.919$, $p < .001$, $\eta_p^2 = .135$). However, the post hoc tests showed that only grade 7 had significantly fewer modifiers per noun phrase than the grades above 7 ($p < .05$) with medium (Cohen's d was .74 for grade 9) to large effect sizes. There were no significant differences among the grades 8, 9, 10, 11, and 12.

The results for the measure of left embeddedness, SYNLE (i.e., the mean number of words before the main verb), showed significant between-group differences, $F(5, 86.184) = 27.363$, $p < .001$, $\eta_p^2 = .358$. The pairwise comparisons showed that the texts for grade 7 had less left embeddedness than grade 8 ($p = .015$, Cohen's $d = .90$), but no significant difference was found between grade 7 and 9 (see Table 6 for descriptive results). The mean for grade 7 was also significantly lower than grades 10, 11 and 12 ($p < .001$) with large effect sizes. While no significant differences were found between grades 8 and 9, both grades had significantly lower mean scores than grades 10, 11 and 12 ($p < .05$) with close-to-large and large effect sizes. The texts for grades 10, 11 and 12 did not differ in terms of the amount of left embeddedness the sentences had. As it can be seen in Figure 4, although there were fluctuations, there was an overall increase in the mean scores for SYNLE from grade 7 to grade 12.

Table 6. Descriptive Results of Three Syntactic Complexity Measures of Coh-Metrix for Reading Texts

Grade Level		SYNLE		SYNSTRUTa		SYNSTRUTt	
	N	M	SD	M	SD	M	SD
7	27	1.45	.71	.15	.04	.15	.05
8	30	2.29	1.10	.15	.04	.14	.03
9	46	1.93	.63	.17	.04	.16	.04
10	31	3.39	1.09	.11	.02	.10	.02
11	47	3.32	1.18	.11	.02	.11	.02
12	31	3.91	1.85	.10	.03	.10	.02

Note. SYNLE = Number of words before the main verb; SYNSTRUTa = Sentence syntax similarity for adjacent sentences; SYNSTRUTt = Sentence syntax similarity across paragraphs; *M* = Mean; *SD* = Standard deviation.

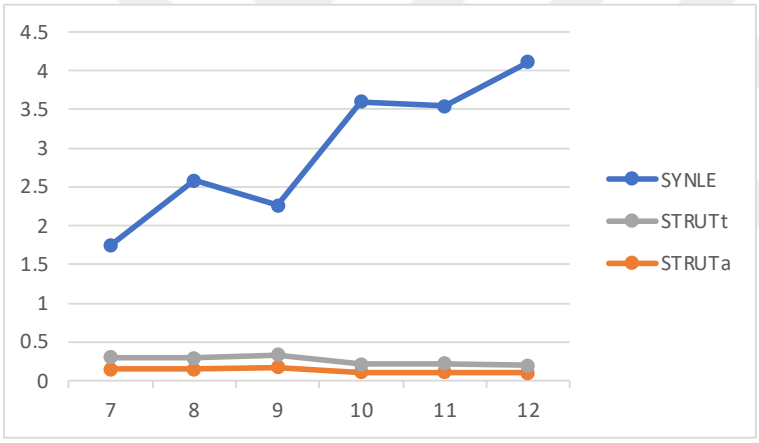

Figure 4. Means of three Coh-Metrix syntactic complexity measures across grades

Two measures of sentence-to-sentence-syntax similarity (Graesser et al., 2014), one of which measured the similarity between the 'adjacent' sentences (i.e., SYNSTRUTa) (op.cit.) and the other measured the parse-tree similarities across the text (i.e., SYNSTRUTt), were used to measure how (dis)similar the texts were in terms of their syntactic structures. Both measures showed significant differences across grades, SYNSTRUTa, $F(5, 88.154) = 16.545$, $p < .001$, $\eta_p^2 = .296$; SYNSTRUTt, $F(5, 87.602) = 23.032$, $p < .001$, $\eta_p^2 = .377$. The post hoc tests for SYNSTRUTa showed that there were no differences between the grade levels 7, 8

and 9 while they had significantly higher mean scores, which means the texts had more similar sentences, than grades 10, 11 and 12 ($p < .05$) with large effect sizes. There were also no significant differences between the grades 10, 11 and 12. The pairwise comparisons for SYNSTRUTt aligned with these results in that no significant differences were found between grades 7, 8 and 9 while they had higher mean scores than grades 10, 11 and 12 ($p < .001$) with large effect sizes (even larger than SYNSTRUTa). As it can be seen in Figure 4, the means for both measures remained stable for the clusters of grades 7, 8 and 9 while there was a decrease in the similarity of the sentences in the texts after grade 9 and stabilizing again between grade 10 and 12.

The results of the tests for the syntactic complexity measures of Coh-Metrix mainly supported the results for the measures of L2SCA and built on them. Overall results for the syntactic complexity of the reading tests showed that, despite some exceptions and mixed results for especially phrase- and clause-based measures, the adjacent grades and clusters of grades (i.e., 5-6(-7), 7-8-9 and 10-11-12) had a tendency to include texts with similar levels of syntactic complexity regarding various measures while each cluster was significantly different from other clusters of grades.

4.1.3 Lexical complexity of the reading texts

Lexical complexity of the reading texts was measured via Lexical Complexity Analyzer (LCA) and Coh-Metrix. Both tools have a word limit for the texts to be analyzed in that Coh-Metrix requires a minimum of 100 words while LCA requires a minimum of 50 words. Therefore, the texts from grades 2, 3, 4, and 6 were not included in either analysis. Grade 5 was included in the analysis of measures from

LCA; however, since the sample did not have sufficient number of texts with more than 100 words, it was excluded from the analysis with measures of Coh-Metrix. Similarly, grade 6 had very few texts with more than 50 words, so it was excluded from the analysis. Since there were different groups and number of texts for either tool, two sets of data were used for lexical complexity measures.

Out of 8 lexical complexity measures, 7 measures showed significant differences between the grades in the one-way ANOVA results while LS-II (i.e., lexical sophistication) showed no significant differences across the grades. In addition, 3 measures showed significant differences between the grades in the ANOVA tests, LD, Welch $F(6, 84.273) = 3.082$, $p = .009$, $\eta_p^2 = .067$; LV, Welch $F(6, 84.467) = 9.526$, $p < .001$, $\eta_p^2 = .158$; $CELEX_{logall}$, Welch $F(5, 87.446) = 5.024$, $p < .001$, $\eta_p^2 = .122$. However, the post hoc tests showed no significant differences between the lowest (i.e., either grade 5 or grade 7 depending on the tool) and the highest grade (i.e., grade 12), and there were few significant differences across the grades.

3 lexical diversity measures (i.e., CTTR, MTLD, SVV1) and one lexical sophistication measure (i.e., $CELEX_{logcontent}$) showed significant differences between the grades in one-way ANOVA tests, CTTR, $F(6, 235) = 51.656$, $p < .001$, $\eta_p^2 = .569$; SVV1, Welch $F(6, 90.995) = 48.747$, $p < .001$, $\eta_p^2 = .510$; MTLD, Welch $F(5, 89.945) = 10.851$, $p < .001$, $\eta_p^2 = .173$; $CELEX_{logcontent}$, $F(5, 206) = 7.962$, $p < .001$, $\eta_p^2 = .162$. Homogeneity of variances was sustained for two measures (i.e., CTTR, $CELEX_{logcontent}$); therefore, Gabriel post hoc test was used to compare the groups while Games-Howell test was used to compare the means of the groups for the measures SVV1 and MTLD (see Table 7 for descriptive results).

Table 7. Descriptive Results of Lexical Diversity Measures for Reading Texts

	LCA				Coh-Metrix		
	CTTR		SVV1		MTLD		
Grade	N	M	SD	M	SD	N	M	SD
5	15	3.80	.61	5.83	3.37	-	-	-
7	37	4.71	.64	9.46	5.02	27	67.93	18.93
8	34	5.26	.69	10.46	4.63	30	69.21	16.34
9	47	5.63	.86	13.59	6.72	46	75.96	24.78
10	31	6.85	.82	23.19	6.93	31	90.28	15.90
11	47	6.57	.79	23.44	7.35	47	91.30	18.99
12	31	6.25	.78	24.76	9.52	31	87.01	25.99

Note. CTTR = Corrected TTR; SVV1 = Squared verb variation; MTLD = Measure of Textual Lexical Diversity; *M* = Mean; *SD* = Standard deviation

The post hoc tests for overall lexical diversity showed somewhat similar results. The pairwise comparisons for the measure of CTTR showed that the texts for grade 5 were significantly less lexically diverse than grade 7 ($p = .002$) and all the above grades ($p < .001$) with large effect sizes (see Table B4 in Appendix B for a complete list of effect sizes for pairwise comparisons). Although no significant differences were found between the texts for grade 7 and grade 8, grade 7 had significantly lower mean than all the grades above 8 ($p < .001$) with large effect sizes as Cohen's *d* ranged from 1.20 (Grade 9) to 2.88 (Grade 10). No differences were found between grade 8 and 9, but the texts for both grades had significantly less lexical diversity than the grades above ($p < .05$) with large effect sizes except that the effect size of the difference between grade 9 and 12 was medium (Cohen's $d = .75$). No significant differences were found between grades 10, 11 and 12. Regarding MTLD measure, no significant differences were found between the grades 7, 8 and 9. However, the texts for grade 7 were significantly less lexically diverse than grade 10 ($p < .001$, Cohen's $d = 1.27$), grade 11 ($p < .001$, Cohen's $d = 1.23$) and grade 12 (*p*

< .025, Cohen's d = .83). The textbook for grade 8 had less lexically diverse texts than grade 10 (p < .001, Cohen's d = 1.30), 11 (p < .001, Cohen's d = 1.24) and grade 12 (p < .027, Cohen's d = .81). There were significant differences between the texts for grade 9 and grade 10 (p = .032, Cohen's d = .68) and 11 (p = .015, Cohen's d = .69), but there was not a significant difference between grade 9 and grade 12. No differences were also found between grade 10, 11 and 12.

As for verb variation measure (i.e., SVV1), the post hoc tests showed no significant differences between grade 5 and 7 although grade 5 had lower mean than all the grades above 7 (p < .001) with large effect sizes. Although no significant differences were found between grade 7 and 8, the texts for grade 7 were less verbally diverse than grade 9 (p = .029) and all the grades above 9 (p = .001) with large effect sizes except that the effect size for the difference with grade 9 was close to medium (Cohen's d = .69). There were no significant differences between grade 8 and 9, but both grades differed significantly from grade 10, 11 and 12 (p < .001) with large effect sizes. No statistically significant differences were found between grades 10, 11 and 12.

Overall, it can be seen that, although there were some exceptions, the lexical diversity measures had a tendency to stabilize between the grades 7, 8 and 9, and then 10, 11 and 12. As it can be seen in Figure 5, despite the clusters of grades, there is a linear increase in the mean scores from grade 5 to grade 10 for both measures of SVV1 and CTTR, whereas there is a drop in the mean scores for CTTR after grade 10. The difference between the results for CTTR and MTLD calls for further examination of the results, though. As for the difference between grade 5 and 7, it can be said that grade 5 had less lexically, but not necessarily verbally, diverse texts

than grade 7, which can be considered as an indication of an increase in the lexical complexity of the texts along with the increase in the CEFR levels of the textbooks.

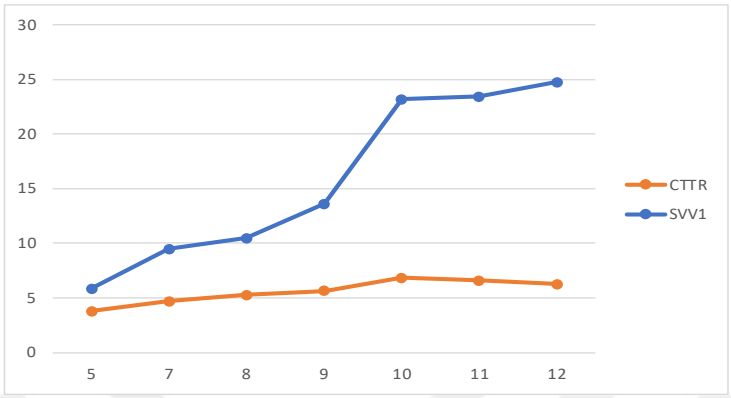

Figure 5. Means of lexical diversity measures of LCA across grade levels

As for the lexical sophistication measure CELEX$_{logcontent}$, the results of Gabriel post hoc test showed that the texts for grade 7 (M = 1.58, SD =.24) had significantly more frequent words, therefore less sophistication, than grade 8 (M = 1.29, SD = .29), 10 (M = 1.23, SD = .19), 11 (M = 1.31, SD = .24) and 12 (M = 1.27, SD = .28) p < .001 with large effect sizes. However, no differences were found between grade 7 and 9 (M = 1.46, SD = .26). For the remaining pairwise comparisons, the only significant difference was between grade 9 (M = 1.46, SD = .26) and 10 (M = 1.23, SD = .19) (p = .009), which had a medium effect size (Cohen's d = .88). The results show that although grade 7 differed from most of the above grades regarding lexical sophistication of the texts, the lexical sophistication of the texts for the remaining grades (i.e., 8, 9, 10, 11 and 12) did not change significantly although grades 10 and 12 had lower mean scores, which indicates higher number of less frequent words.

Overall, the results of lexical complexity measures aligned with the results of syntactic complexity measures to some extent in that similar cluster of levels did not show significant differences in lexical complexity except for some measures.

4.2 Results for research question 2

The second research question investigated the linguistic complexity (i.e., syntactic complexity and lexical complexity) of the listening texts across grades. Because of the few numbers of written texts in the textbooks for grade 2, 3 and 4, these grades could not be included in the analysis of the linguistic complexity of the reading texts. However, in order to be able to compare the texts in these textbooks, the second research question investigated the linguistic complexity of the listening texts, which formed the main input of the textbooks for the primary school. The tools L2SCA and LCA were used to measure linguistic complexity of the audio texts.

4.2.1 Syntactic complexity of the listening texts

A total of 6 measures were used to investigate syntactic complexity of the listening texts. 3 measures of syntactic complexity met the conditions for normality, as the skewness and kurtosis values were within the range after removal of some extreme outliers and winsorizing. Since none of these measures met the Levene's homogeneity of variances assumption, Welch F is used to report the results, and Games-Howell test was used for post hoc comparisons.

The three length-based measures showed significant differences between the grades, MLS, $F(10, 108.915) = 50.397$, $p < .001$, $\eta_p^2 = .584$; MLT, $F(10, 110.126) = 50.638$, $p < .001$, $\eta_p^2 = .585$; MLC, $F(10, 111.950) = 48.245$, $p < .001$, $\eta_p^2 = .461$. The post hoc tests for MLS showed that the texts for grade 2 were significantly

shorter than all the grades above ($p < .05$) with large effect sizes except that Cohen's d was .75 for the difference with grade 3 and .81 for the difference with grade 4 (see Table 8 for descriptive results; see Table B5 in Appendix B for a complete list of effect sizes for pairwise comparisons). There were no significant differences between grades 2, 3 and 4 for the measure MLT. For MLC, there were no significant differences between grade 2 and 3, but a significant difference was found between grade 2 and 4 ($p = .010$, Cohen's $d = .71$). It can be claimed that there was a slight increase in length-based measures in the texts over grades 2 to 4. When the medium effect size is considered, it can be claimed that these grades are more similar to each other than the grades above them as they had significantly lower means than grade 5 and all the grades above 5 for both MLT, MLS ($p < .001$) and MLC measures ($p < .05$). The effect sizes of the differences were close to large or large.

The results for length-based measures showed variances across grades 5 to 12 in that although grade 5 had texts with significantly lower mean length of T-unit than the grades above 6 ($p < .05$) with large effect sizes, no significant differences were found between grade 5 and grades 6, 8 and 9 for MLS measure. Similarly, there were no significant differences between grade 6 and grades 8-9-10 for MLT (however, the large effect size for the difference with grade 10 requires further investigation), and grade 6 and 8-9 for MLS. As for grade 7, it had a significantly higher mean than all the grades below it for both MLS and MLT while the only significant difference with grade 7 and the above grades was with grade 9 for MLS, which had a lower mean score with a large effect size (Cohen's $d = 1.32$). It can be seen that grade 7 had longer sentences and T-units than the grades below it and longer sentences than grade 9 while no significant differences were found with other grades. However, MLC measures showed a somewhat different trend in that although there were no

significant differences between grades 10, 11 and 12, grade 11 had significantly longer clauses than all the other grades below ($p < .01$) except grade 7. No other significant differences were observed among grades 5 to 12 in terms of MLC.

Table 8. Descriptive Results of Syntactic Complexity Measures for Listening Texts

Grade	MLS		MLT		MLC	
	M	SD	M	SD	M	SD
2	3.46	.95	3.88	.88	3.96	.84
3	4.28	1.21	4.45	1.17	4.52	1.20
4	4.30	1.10	4.37	.90	4.66	1.10
5	5.89	1.32	5.57	.85	5.87	1.51
6	6.03	1.30	6.24	1.34	5.81	1.27
7	8.61	2.30	8.36	2.26	7.45	2.07
8	6.92	2.61	7.16	2.03	6.21	1.49
9	6.21	1.14	6.81	1.21	6.47	.97
10	9.35	3.65	9.12	3.57	7.51	2.21
11	9.94	1.81	9.76	1.68	7.34	.63
12	9.73	3.45	9.72	3.59	7.54	2.30

Note. MLS = Mean length of sentence; MLT = Mean length of T-unit; MLC = Mean length of clause; *M* = Mean; *SD* = Standard deviation

The overall results may indicate that the lower-level grades 2-3-4 tended to have texts with similar length of sentence and/or clauses while they had texts with significantly shorter sentences than the texts for above grades. The texts for grade 7 and grade 11 stood out among the other texts below grade 10 in that they had relatively longer sentences and clauses, which can be seen in Figure 6 as well.

3 measures did not meet the assumptions of normality, as there were too many outliers especially for the lower grades and the skewness and kurtosis values were not within range; therefore, nonparametric tests (i.e., Kruskal-Wallis and Spearman's correlation coefficient) were used to compare the means of the grades for the measures of subordination (i.e., DC/T), sentence coordination (i.e., T/S) and noun phrase complexity (i.e., CN/C). The results of the Kruskal-Wallis nonparametric test showed significant differences between the grades for all three

measures, DC/T, $H(10) = 253.252$, $p < .001$; T/S, $H(10) = 41.817$, $p < .001$; CN/C, $H(10) = 253.252$, $p < .001$.

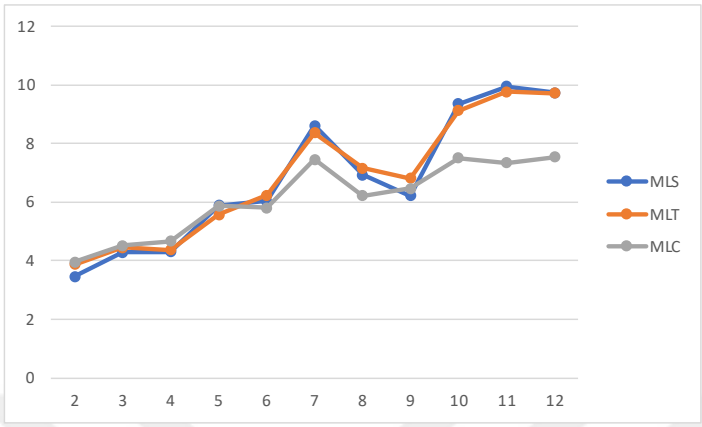

Figure 6. Means of length-based syntactic complexity measures for listening texts across grade levels

As for the post hoc tests, 55 independent samples tests needed to be conducted to analyze the difference between the grades; however, this would inflate the Type I error substantially. Therefore, Spearman's correlation coefficient (Spearman, 1910) was used to identify if there was a significant relationship between the grade levels and the measures. The Spearman's correlation tests showed that there was a strong, positive correlation between grade levels and the syntactic complexity measure DC/T, $r_s(361) = .767$, $p < .01$; a medium correlation with CN/C $r_s(361) = .544$, $p < .01$ and a weak correlation with T/S, $r_s(361) = .234$, $p < .01$.

4.2.2 Lexical complexity of the listening texts

5 measures of lexical complexity (i.e., LD, LS-II, CTTR, SVV1 and LV) were used to investigate the lexical complexity of the listening texts. The analyses included the texts starting from grade 4 as there were very limited number of texts with more than 50 words in the textbooks for grades 2 and 3. 4 out of 5 measures showed significant

between-group differences in the one-way ANOVA tests. The result of the one-way ANOVA test was not significant for the lexical sophistication measure LS-II. Although there were significant results for one-way ANOVAs for the measures of lexical variation (i.e., LV) (Welch $F(8, 70.348) = 7.497$, $p < .001$, $\eta_p^2 = .265$) and lexical density (i.e., LD) (Welch $F(8, 70.929) = 2.134$, $p = .043$, $\eta_p^2 = .072$), the post hoc tests did not show meaningful differences between the grades for LV as there were no significant differences between the low (e.g., 5) and the high grades (e.g., 11, 12). No significant differences were found in the post hoc test for LD. Therefore, the results are not further presented and discussed.

For the remaining two lexical diversity measures (i.e., CTTR and SVV1), Levene's homogeneity of variances was not sustained; therefore, Welch's F is used to report the results. The results showed significant between-group differences, CTTR, $F(8, 70.927) = 90.954$, $p < .001$, $\eta_p^2 = .711$; SVV1, $F(8, 72.918) = 39.533$, $p < .001$, $\eta_p^2 = .633$. Since the homogeneity of variances was not assumed, Games-Howell post hoc test was used to compare different grades.

The results of the pairwise comparisons for CTTR showed no significant differences between grade 4 and 5 while both grades had significantly lower means than all the above grades ($p < .05$) with large effect sizes (see Table 9 for descriptive results, see Table B5 in Appendix B for a complete list of effect sizes for pairwise comparisons). There were not any significant differences between grades 6, 7 and 8 as well, whereas they differed significantly from all the grades above 8 ($p < .05$) with large effect sizes. No significant differences were also found among the grades 9, 10 and 12. However, grade 11 had the highest mean among all the other grades ($p < .001$) and the effect size was large for all the pairwise comparisons. In other words, although there were clusters of grades that had similar results for lexical diversity as

in the case of grades 4 and 5; 6, 7 and 8; 9, 10 and, with an exception, grade 12, grade 11 had the most lexically diverse listening texts.

Table 9. Descriptive Results of Lexical Diversity Measures for Listening Texts

		CTTR		SVV1	
Grade	N	M	SD	M	SD
4	15	2.62	.72	2.92	2.06
5	17	3.22	.59	4.13	2.25
6	28	4.03	.71	7.09	4.36
7	25	4.29	.38	8.69	3.21
8	35	4.28	.78	8.31	4.45
9	21	5.42	.67	12.89	5.78
10	17	5.30	.76	14.43	5.19
11	21	6.51	.39	25.39	6.61
12	18	5.18	.84	14.30	6.62

Note. CTTR = Corrected TTR; SVV1 = Squared verb variation; MTLD = Measure of Textual Lexical Diversity; *M* = Mean; *SD* = Standard deviation

The results of the post hoc tests for the verb diversity measure SVV1 showed similar results with a few exceptions. No significant differences were found between grade 4 and 5, but grade 4 had a significantly lower mean than all the above grades ($p < .05$) with large effect size. Although grade 5 and 6 were not significantly different, the texts for grade 5 had significantly lower scores than all the grades above 6 ($p < .001$) with large effect size. No differences were found between grades 6, 7 and 8, but grade 6 had a significantly lower mean score from all the grades above 8 ($p < .05$) with large effect size. Similarly, no differences were found between grades 7, 8 and 9, but grade 8 had significantly lower mean than all the grades above 9 ($p <.05$) with large effect sizes while grade 7 had a lower mean than grades 10 and 11 ($p < .05$) with large effect sizes. Although no differences were found between grade 7 and 12 in terms of the diversity of verbs the texts had, the effect size was quite large (Cohen's $d = 1.07$), which may be an indication of the effect of small sample size, so it requires further investigation. Similar to the results

of CTTR, no differences were found between grades 9, 10 and 12, but grade 11 was significantly different from all the other grades ($p < .001$) with large effect sizes.

The results of the lexical complexity analyses showed a tendency for the adjacent grades to have nonsignificant differences, so similar levels of lexical diversity. As it can be seen in the Figure 7, the mean scores mainly had a linear increase especially for the grades below 10. However, grade 11 stood out from all the other grades in that it had the most lexically complex/diverse texts. Contrary to expectations, there were no significant differences for either between grade 9, (CEFR A1/A2), 10 (CEFR A2, B1) and grade 12, which was regarded as CEFR B2+.

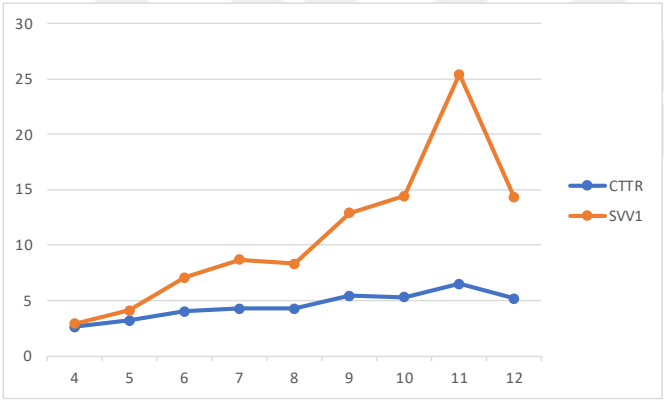

Figure 7. Means of lexical diversity measures for listening texts across grade levels

The overall results for the analysis of listening texts showed that there were significant differences between lower grade levels (e.g., grade 2-3-4 and 5-6-7) in terms their sentence-length while there were more mixed results as the grade levels increased. However, the lexical diversity results showed that, although there were some exceptions, there were clusters of grades 4-5, 5-6, 6-7-8, 7-8-9, 9-10-12 that had similar lexically diverse texts while grade 11 had significantly the most lexically diverse texts.

4.3 Results for research question 3

The third research question aimed to measure how similar and/or different the textbooks used in grade levels 11(CEFR B1+/B2) and 12 (CEFR B2+) in state schools are to two internationally published textbooks with similar CEFR levels, namely Interchange3 (CEFR B1/B1+) and Passages1 (CEFR B2). The textbooks were analyzed by the same measures used to compare the L2 English textbooks used in the state schools. Overall, 3 readability formulas, 12 syntactic complexity measures and 8 lexical complexity measures were used to measure and compare the linguistic complexity of the textbooks.

4.3.1 Comparison of reading texts in terms of readability levels

The results of Levene's test of homogeneity of variances was not sustained for the traditional readability formulas; therefore, Welch *F* and the results of Games-Howell post hoc test are used to report the results. The results showed that there were significant differences between the locally and internationally published textbooks, Flesch Ease, $F(3, 58.087) = 7.211$, $p < .001$, $\eta_p^2 = .150$; Flesch-Grade, $F(3, 57.962) = 10.357$, $p < .001$, $\eta_p^2 = .185$. The results of Flesch Ease formula showed that grade 11 ($M = 72.61$, $SD = 9.00$) had significantly more readable texts than Interchange3 ($M = 65.45$, $SD = 10.86$) ($p = .018$) and Passages1 ($M = 64.79$, SD = 9.11) ($p = .009$) with medium effect sizes. Similarly, the results for Flesch-Kincaid Grade Level formula showed that the texts for grade 11 ($M = 6.52$, $SD = 1.71$) were more readable than the texts in Interchange3 ($M = 8.58$, $SD = 2.13$) ($p < .001$, Cohen's $d = 1.06$) and Passages1 ($M = 8.21$, $SD = 1.74$) ($p = .003$, Cohen's $d = .97$). No significant differences were found between the textbook for Grade 12 and the internationally published textbooks.

As for Coh-Metrix L2 Readability Index, although the results of the one-way ANOVA showed significant differences between the textbooks, $F(3, 121) = 2.968$, $p = .035$, $\eta_p^2 = .069$, the results of Gabriel post hoc test did not show any significant differences among the textbooks.

4.3.2 Comparison of reading texts in terms of syntactic complexity

In order to investigate the similarities of the textbooks regarding their syntactic complexity, 12 syntactic complexity measures (i.e., READSL/MLS, MLT, MLC, DC/T, T/S, CN/C, SYNLE, SYNNP, SYNSTRUTa, SYNSTRUTt, $SYNMED_{word}$, $SYNMED_{pos}$) were used. Two measures of overall sentence length were included as there were slight differences between the results of the written texts in the textbooks for the state schools in terms of these measures. However, the results of the one-way ANOVAs and the post hoc tests showed that the results for both measures were the same; therefore, only one overall sentence length measure (i.e., READSL) is reported. The results of a series of one-way ANOVAs showed that 8 out of 11 measures showed significant differences between the texts in the textbooks for the state schools and the internationally published textbooks. However, the textbooks did not differ in terms of the measures $SYNMED_{pos}$, $SYNMED_{wrd}$, CN/C and MLC.

Two length-based measures of syntactic complexity (i.e., READSL, MLT) showed significant differences among the groups. Levene's homogeneity of variances was not sustained; therefore, Welch *F* is reported, and Games-Howell post hoc test was conducted to analyze the differences between the groups. One-way ANOVAs showed significant differences between the textbooks for each length-based measure, READSL, $F(3, 60.935) = 15.802$, $p < .001$, $\eta_p^2 = .194$; MLT, $F(3, 58.916) = 6.229$, $p = .001$, $\eta_p^2 = .089$. In terms of sentence length (i.e., READSL),

the texts for grade 11 had significantly shorter sentences than Interchange3 ($p < .001$, Cohen's $d = 1.58$) and Passages1 ($p = .032$, Cohen's $d = .88$) (see Table 10 for descriptive results; see Table B6 and B7 in Appendix B for a complete list of effect sizes for pairwise comparisons). Similarly, the post hoc tests for the measure of MLT showed that that there were significant differences between the texts for grade 11 and Interchange3 ($p = .007$) and Passages1 (p = .016) with medium effect size as the Cohen's d was .81 for Interchange3 and .82 for Passages1. No differences were found between grade 12 and the internationally published textbooks for either measure. For MLC measure, no significant differences were found between the textbooks for grades 11 and 12 and the internationally published textbooks. The results show that the texts for grade 11 had significantly shorter sentences and T-units than the texts in the internationally published textbooks of a similar CEFR level with medium to large effect size.

Table 10. Descriptive Results of Four Measures of Syntactic Complexity for Comparing Reading Texts

Grade		READSL		MLT		SYNSTRUTa		SYNSTRUTt	
	N	M	SD	M	SD	M	SD	M	SD
11	47	13.69	2.52	12.94	1.95	.11	.02	.10	.02
12	31	15.46	4.93	14.66	4.30	.10	.03	.10	.02
Int3	32	17.92	2.81	15.13	3.26	.08	.01	.08	.01
Pass1	23	16.06	2.84	14.69	2.30	.09	.02	.08	.02

Note. READSL = Sentence length; MLT = Mean length of T-unit; SYNSTRUTa = Sentence syntax similarity for adjacent sentences; SYNSTRUTt = Sentence syntax similarity across paragraphs; *M* = Mean; *SD* = Standard deviation.

As for the measures of sentence (dis)similarity, the data for SYNTRUTa did not meet Levene's homogeneity of variances assumption; therefore, Welch's F is

reported, and Games-Howell test was used for pairwise comparisons. However, since the measure SYNSTRUTt met the assumption, ANOVA's *F* is used to report the results, and Gabriel post hoc test was used for the comparisons. The results showed significant differences between the textbooks, SYNSTRUTa, $F(3, 62.654) = 11.820$, $p < .001$, $\eta_p^2 = .163$; SYNSTRUTt, $F(3, 129) = 10.097$, $p < .001$, $\eta_p^2 = .190$. Regarding the similarity of adjacent sentences (i.e., SYNSTRUTa), the texts for grade 11 had more similar sentences than Interchange3 ($p < .001$) and Passages1 ($p = .020$) with large and medium effect size, respectively. In terms of similarity across the texts (i.e., SYNSTRUTt), grade 11 had a significantly higher mean than both Interchange3 ($p < .001$, Cohen's $d = 1.24$) and Passages1 ($p = .005$; Cohen's $d = .87$). As for grade 12, there were no differences between grade 12 and Passages1 in terms of either measure. However, Interchange3 had significantly lower mean scores for both the measure SYNSTRUTa ($p = .014$, Cohen's $d = .80$) and SYNSTRUTt ($p = .001$, Cohen's $d = .94$). The results indicate that the textbook for grade 11 had texts with more similar sentence structures than the internationally published textbooks, while grade 12 had texts that had more similar sentences than Interchange3.

The measures for left embeddedness (i.e., SYNLE), sentence coordination (i.e., T/S) and noun phrase complexity (i.e., SYNNP) did not meet the assumptions for Levene's homogeneity of variances; therefore, Welch's *F* is used to report the results. On a similar vein, Games-Howell post hoc test was used for pairwise comparisons. The results showed significant differences between the textbooks for each measure; SYNLE, $F(3, 64.584) = 2.780$, $p = .048$, $\eta_p^2 = .055$; SYNNP, $F(3, 68.089) = 3.426$, $p = .022$, $\eta_p^2 = .063$; T/S, $F(3, 59.648) = 5.858$, $p < .001$, $\eta_p^2 = .176$. In terms of left embeddedness, the only significant difference was that grade 11 ($M = 3.32$, $SD = 1.18$) had sentences with fewer number of words before the main verb

than Interchange3 (M = 4.12, SD = .94) (p = .034), but the effect size was close to medium (Cohen's d = .64). Although the results for SYNNP was significant, the post hoc tests showed that only the textbook for grade 11 (M = .86, SD = .19) had texts with fewer number of modifiers for noun phrases than Interchange3 (M = .75, SD = .16) and Passages1 (M = .76, SD = .10) (p = .036); however, the effect size was close to medium for both results (Cohen's d = .61 for Interchange3; Cohen's d = .67 for Passages1), which indicates a need for further investigation. As for the measure T/S, Interchange3 (M = 1.21, SD = .12) had higher amount of sentence coordination than the textbook for grade 11 (M = 1.11, SD = .06) (p = .002, Cohen's d =.92) and grade 12 (M = 1.09, SD = .08) (p = .001, Cohen's d = 1.01).

Since Levene's homogeneity of variances was sustained for DC/T, ANOVA's F is used for reporting the results, and Gabriel was used for pairwise comparisons. The results showed significant differences between the groups, $F(3, 127) = 12.392$, $p < .001$, $\eta_p^2 = .226$. The pairwise comparisons showed that both internationally published textbooks had higher subordination than the high school textbooks. Interchange3 (M = .63, SD = .05) had texts with higher amount of subordination than the textbook for grade 11 (M = .45, SD = .17) ($p < .001$, Cohen's d = 1.16) and grade 12 (M =.43, SD = .20) ($p < .001$, Cohen's d = 1.16). Similarly, Passages1 (M = .63, SD = .24) had higher amount of subordination than the textbook for grade 11 (p = .002, Cohen's d = .88) and grade 12 (p = .003, Cohen's d = .91).

The results mainly showed that the texts for grade 11 were syntactically less complex than Interchange3 and Passages1 in terms of overall sentence length, similarity across sentences, subordination and sentence coordination. Although it had higher amount of noun phrase complexity than both international textbooks, the small effect size and the nonsignificant results of the measure of CN/C indicated a

need for further examination. As for the textbooks for grade 12, the texts were similar to Passages1 for the majority of the measures (except for the higher amount of subordination for Passages1). However, for certain syntactic complexity measures, Interchange3 also had more complex sentences with medium to large effect size than the textbook for grade 12.

4.3.3 Comparison of reading texts in terms of lexical complexity

Out of 8 lexical complexity measures, the results for 5 measures showed significant differences between groups while there were no differences between the texts of different books for the measures of verb diversity (i.e., SVV1), lexical diversity measure of MTLD and lexical sophistication measure of $CELEX_{logcontent}$.

The measure for lexical density (i.e., LD) showed significant differences between groups, Welch $F(3, 64.009) = 3.493$, $p = .021$, $\eta_p^2 = .062$. The post hoc test showed a significant difference between the textbook for grade 11 ($M = .50$, $SD = .03$) and Passages1 ($M = .52$, $SD = .02$) with medium effect size (see Table B8 in Appendix B for a complete list of effect sizes for comparing the listening texts in the textbooks). No other significant differences were found.

Only one measure of lexical diversity, CTTR, showed significant differences between groups, Welch $F(3, 64.926) = 4.938$, $p = .004$, $\eta_p^2 = .075$. No significant differences were found between the textbook for grade 11 and Interchange3 and Passages1, which indicated that the texts had similar levels of lexical diversity. However, Passages1 ($M = 6.81$, $SD = .49$) had more lexically diverse texts than the textbook for grade 12 ($M = 6.25$, $SD = .78$) with a medium effect size (Cohen's $d = .84$). No differences were found between the texts for grade 12 and Interchange3. Although CTTR showed significant differences (though with medium effect size), no

significant difference was found between the textbooks for the measure MTLD, which might be related to the way each measure treats and measures lexical diversity. Regarding verb diversity, no significant differences were found among the textbooks.

As for lexical sophistication measures, the results of one-way ANOVAs showed that two lexical sophistication measures had significant differences between groups, LS-II, $F(3, 129) = 5.949$, $p = .001$, $\eta_p^2 = .122$; $CELEX_{logall}$, Welch $F(3, 61.252) = 10.524$, $p < .001$, $\eta_p^2 = .147$. Regarding the measure of LS-II, both grade 11 and grade 12 had texts with higher lexical sophistication than Interchange3 with an effect size close to large for grade 11 (Cohen's $d = .94$) and medium for grade 12 (Cohen's $d = .77$) (see Table 11 for descriptive results). No significant differences were found between Passages1 and the locally published textbooks in terms of LS-II. For $CELEX_{logall}$, Games-Howell post hoc test showed no differences between Passages1 and the textbooks used in high schools. However, the texts for grade 12 had significantly higher amount of less frequent words than Interchange3 ($p = .001$) with large effect size. Although there was a significant difference between grade 11 and Interchange3 ($p = .015$), the effect size was close to medium (Cohen's $d = .67$).

Table 11. Descriptive Results of Lexical Sophistication Measures for Comparing Reading Texts

	$CELEX_{logall}$		LS-II	
Grade	M	SD	M	SD
11	3.04	.08	.24	.06
12	3.01	.10	.23	.05
Int3	3.09	.05	.19	.04
Pass1	3.00	.08	.23	.04

Note. $CELEXl_{ogall}$ = CELEX Log frequency for all words; LS-II = Lexical sophistication; *M* = Mean; *SD* = Standard deviation

As for lexical variation, the results of ANOVA showed significant differences between the groups LV, $F(3, 129) = 6.281$, $p = .001$, $\eta_p^2 = .127$. However, only grade 11 ($M = .74$, $SD = .06$) had higher lexical variation than Interchange3 ($M = .68$, $SD = .06$) ($p < .001$, Cohen's $d = .99$).

Combined with the effect sizes, the results indicate that the textbooks used in the state schools were similar with the internationally published textbooks in terms of the diversity of lexical words used. As for lexical sophistication, the textbooks for grades 11 and 12 tended to have texts with words that are less frequent than the texts in Interchange3; however, the high school texts were similar to Passages1 in terms of lexical sophistication of the texts.

4.3.4 Comparison of listening texts for syntactic complexity

6 measures of syntactic complexity by L2SCA (i.e., MLS, MLT, MLC, T/S, DC/T, CN/C) were used to analyze the difference between the listening texts. The results of the one-way ANOVAs showed that there were no significant differences between the textbooks for the measures MLS, MLT and T/S. In terms of subordination, one-way ANOVA results showed that there were significant differences between the textbooks for DC/T, $F(3, 108) = 2.872$, $p = .040$, $\eta_p^2 = .074$. However, the Gabriel post hoc test showed no significant differences between the groups.

As for the remaining measures MLC and CN/C, Levene's homogeneity of variances was not sustained; therefore, Welch's F is used to report the result of ANOVA, and Games-Howell test was used to compare the groups. Both measures had significant results MLC, $F(3, 46.554) = 8.938$, $p < .001$, $\eta_p^2 = .135$; CN/C, $F(3, 46.554) = 4.631$, $p = .006$, $\eta_p^2 = .081$. Pairwise comparisons for MLC showed that there were no significant differences between the texts for grade 12 and the

internationally published textbooks. However, the listening texts for grade 11 (M = 7.34, SD = .63) had longer clauses than both Interchange3 (M = 6.41, SD = .79) (p < .001, Cohen's d = 1.29) and Passages1 (M = 6.68, SD = .82) (p < .05, Cohen's d = .89). Regarding the number of complex nominals, the only significant difference was that the texts for grade 11 (M = .64, SD = .16) had higher number of complex nominals per clause than Interchange3 (M = .48, SD = .17) (p = .004), and the effect size was close to large (Cohen's d = .96).

The results showed that the listening texts for grade 12 were similar to the international textbooks in terms of their syntactic complexity. The listening texts for grade 11 were also similar to the international textbooks in terms of sentence length and amount of sentence coordination, but they had longer clauses than both textbooks and higher noun phrase complexity than Interchange3.

4.3.5 Comparison of listening texts in terms of lexical complexity

5 measures of lexical diversity (i.e., LD, LS-II, CTTR, SVV1 and LV) by LCA were used to analyze and compare the lexical complexity of the listening texts in the textbooks. The one-way ANOVA for the measure of lexical density (i.e., LD) did not show significant results. 2 of the remaining measures (i.e., CTTR and LV) violated Levene's homogeneity of variances assumption; therefore, Welch's F is used to report the results, and Games-Howell test was used for pairwise comparisons. For the measures LS-II and SVV1, ANOVA's F and Gabriel post hoc test were used. All four measures showed significant between-group differences, LS-II, $F(3, 103) = 33.915$, $p < .001$, $\eta_p^2 = .497$; CTTR, $F(3, 49.337) = 32.560$, $p < .001$, $\eta_p^2 = .373$; SVV1, $F(3, 108) = 12.194$, $p < .001$, $\eta_p^2 = .253$; LV, $F(3, 50.691) = 22.644$, $p < .001$, $\eta_p^2 = .349$.

The results for lexical sophistication measure showed that the texts for grade 11 (M = .25, SD = .03) and grade 12 (M = .22, SD = .05) had significantly less frequent words than Interchange3 (M = .13, SD = .04) and Passages1 (M = .16, SD = .04) (p < .001), and the effect size was large (see Table B9 in Appendix for a complete list of effect sizes of lexical complexity measures for comparing the listening texts in the textbooks).

As for lexical diversity measure (i.e., CTTR), the results of the post hoc tests showed that the texts for grade 12 (M = 5.18, SD = .84) were less lexically diverse than the texts in Passages1 (M= 6.11, SD = .77) (p = .003, Cohen's d = 1.15) while no differences were found with Interchange3 (M = 5.38, SD = .57). As for the textbook for grade 11 (M= 6.51, SD = .39), it had texts that were more lexically diverse than the texts in Interchange3 (M = 5.38, SD = .57) (p < .001) with a large effect size (Cohen's d = 2.31) while no differences were found with Passages1. The results for verb diversity measure (i.e., SVV1) aligned with the results of CTTR. The textbook for grade 11 (M = 25.39, SD = 6.61) had texts that were more verbally diverse than Interchange3 (M = 17.25, SD = 5.33) (p = < .001, Cohen's d = 1.35) while there was no significant difference with Passages1 (M = 21.95, SD = 8.02). The textbook for grade 12 (M = 14.30, SD = 6.62) had texts that were less verbally diverse than Passages1 (M = 21.95, SD = 8.02) (p = .001, Cohen's d = 1.03) while there were no significant differences with Interchange3 (M = 17.25, SD = 5.33).

The results for LV showed significant differences between the texts for grade 11 (M = .73, SD = .05) and Passages1 (M = .61, SD = .06) (p < .001, Cohen's d = 1.99). The texts for grade 12 were significantly more lexically diverse than both Interchange3 (M = .69, SD = .09) (p = .003, Cohen's d = .84) and Passages1 (M = .61, SD = .06) (p < .001, Cohen's d = 1.95).

CHAPTER 5

DISCUSSION

This chapter starts with a discussion of the results obtained from the analyses in relation to the research in the literature on SLA theories. Based on the findings and discussion, it further presents several implications for the improvement of the textbooks used in the state schools and the use of lexical complexity measures (combined with readability formulas) in second language education research. In the final part, the limitations of the present study are presented, and the suggestions for future research are discussed.

5.1 Discussion

5.1.1 Text difficulty of the reading texts in locally published textbooks

The present study investigated the complexity/difficulty of the reading texts in the L2 English textbooks used in state schools in Turkey in several aspects through readability formulas and linguistic complexity measures. It was expected that there would be an increase in the linguistic complexity of the reading texts along with the increase in the grade levels, more specifically CEFR levels attributed to each grade and/or clusters of grades. The results partially supported the initial hypotheses.

The results of the readability formulas showed both similar and different results in terms of the readability of the texts in different grades. The results of both Flesch-Kincaid Grade Level and Coh-Metrix L2 Readability Index showed that the textbook for grade 9 (i.e., CEFR A1/A2) was more readable than the textbooks for the higher grades, which ranged from A2+/B1 to B2+. Although grade 9 and grade 11 were rated similar in terms of the readability levels of the texts they had by Flesch

Ease Formula, considering the weak power of the formula in differentiating among the texts of various levels (Crossley et al., 2011), it can be claimed that the result may be related to the formula itself. Similarly, Crossley et al. (2011) further indicated the superiority of Coh-Metrix L2 Readability Index over the traditional formulas, which was also observed in the present study in that the practical significance of the difference between grade 9 and the other grades was mainly higher for Coh-Metrix L2 Readability Index than the traditional readability formulas. As for the grade levels 10, 11 and 12, the traditional readability formulas showed that the texts for grade 11 were more readable than the texts for grade 10 (although the effect size was close to medium) and less readable than grade 12. The results were in line with the findings of the study by Morales (2019) to a degree in that no significant differences were found between grade 10 and 12, while grade 11 had easier texts to read. However, the readability of the texts was not different for Coh-Metrix L2 Readability Index, which does not only focus on surface levels features of language, but also on cohesion and word frequency, and this may affect the readability of a text. In addition, the results of the Coh-Metrix L2 Readability Index aligned with the results of the linguistic measures used for the analysis of the texts. Although readability formulas have been widely criticized for their unidimensional approach to the texts, the results of the readability formulas used in the present study, especially the results of Coh-Metrix L2 Readability Index, supported the results of more fine-grained indices and measures of text difficulty (i.e., linguistic complexity measures). This can be a further support to the findings by Vajjala & Meurers (2012) that although readability formulas have relatively weak power to differentiate between the levels, the accuracy of the grading of the texts is higher when they are combined with linguistic complexity measures.

When the syntactic complexity of the texts for different grade levels is considered, the results for the measures showed significant differences among the grade levels with medium to large effect sizes. Further analysis of the significant results showed that the adjacent grade levels or a cluster of grade levels tended to have similar levels of syntactic complexity, especially in terms of overall sentence-length measures (i.e., MLS, MLT) and (dis)similarity of the texts. The textbooks for grade levels 5-6 (and for some measures grade 7), which were reported to correspond to CEFR A1 level, had texts that were similar in terms of length-based measures. Similarly, the textbooks for grades 7, 8 and 9 tended to have similar results for the overall sentence-length measures, although there were some exceptions with medium effect size. For almost all syntactic complexity measures, the textbooks for grades 10, 11 and 12 had nonsignificant results, and therefore it can be claimed that they had texts with similar syntactic complexity levels. The clustering of textbooks for different grades was also prominent in text similarity measures (i.e., SYNSTRUTa, SYNSTRUTt) in that the texts for grade levels 7, 8 and 9 had more similar sentences, which might result in easier and less cognitively demanding texts, than the textbooks for grades 10, 11 and 12. These are in line with the results of the study by Jin et al. (2020) in that they observed a stabilization of the syntactic complexity measures among the clusters of grades especially in the lower levels (i.e., grade 1-4) and then in higher levels (i.e., grades 10-12). They further argued that all dimensions of syntactic complexity shouldn't be expected to show a linear increase as the grade levels increase, but the materials developers should rather focus on various dimensions of syntactic complexity in the texts.

Although the present study and the results of the study by Jin et al. (2020) show similar results, it should be noted that there were mixed results for some

measures of subdimensions of syntactic complexity in the present study. The results for the measures of subordination (i.e., DC/T) and left embeddedness (i.e., SYNLE) were in line with the results of overall sentence-length measures in that the clusters of grade levels had texts with similar levels of subordination and embeddedness (i.e., grades 5-6, 7-8-9, 10-11-12) although there were some exceptions of significant differences with moderate effect sizes between the grades in the same clusters. However, the results for the measures of sentence coordination (i.e., T/S), and phrasal and clausal complexity (i.e., CN/C, MLC, SYNNP) were not always significantly different across the grades. To exemplify, the measure for sentence coordination (i.e., T/S) showed no significant differences between grade 5 and all the grades above. On a similar vein, the texts for grade 5 and 6 had significantly fewer complex nominals per clause than the grades above 6, and the texts for grade 7 had fewer complex nominals per clause than the grades above the cluster of 7-8-9. However, no significant differences were found between grade 8 and all the above grades in terms of CN/C, and grade 9 was only significantly different from grade 10, but not grade 11 and 12 in terms of CN/C. In other words, the texts above 7 did not necessarily significantly increase in noun phrase complexity.

The overall analysis of the results for syntactic complexity measures showed that there were clusters of grades that stabilized in syntactic complexity for many of the measures. When the clusters of 5-6 and 7-8-9 are considered, it can be said that they match the CEFR levels of A1 and A2, respectively, so nonsignificant results between these grades were expected. Similarly, the texts for grade 10 (CEFT A2+/B1) were significantly more complex in many measures (e.g., sentence-length, left embeddedness and similarity of the texts) than grade 9, which can be an indication of the transition from CEFR A levels to B levels. The results of the

previous studies also showed a stabilization of syntactic complexity measures among the grades 10-12 (Jin et al., 2020); however, considering that each grade level corresponds to a different CEFR proficiency level in the textbooks used in Turkey (i.e., grade 10 = A2/B1; grade 11 = B1+/B2; grade 12 = B2+), an increase in at least one aspect of linguistic complexity is expected, which is the third part of the research question, namely the lexical complexity of the texts.

Considering that lexical features like diversity and sophistication are significant constructs that affect the comprehensibility and/or difficulty of the texts for learners (Crossley, Allen & McNamara, 2012), it was expected that the lexical complexity of the texts would show a linear increase with the grade level. However, the lexical complexity measures for the reading texts had mixed results.

One lexical diversity measure, SVV1, aligned with the syntactic complexity measures to a great extent by forming a cluster of levels between grades 5-7 (as grade 6 had too few texts over 50 words to be included in the analysis), 7-8-9 and 10-11-12. Similarly, the results for CTTR and MTLD showed that there were no significant differences regarding lexical diversity of the texts for grade 10-11-12, which was contrary to expectations in that the textbooks did not necessarily increase in lexical diversity. In addition, there was a significant difference between grade 7 and grade 9 in terms of CTTR results and grade 9 had more lexically diverse texts although there were no significant differences between grade 8 and 9. MTLD also did not show any significant differences between grades 7, 8 and 9. When we consider that although grade 9 has texts for A1 level, it is expected to "be more advanced in terms of some vocabulary and structures compared to 2nd-8th Grade English Curriculum so that students can also receive new input while A1 level in 2 they are revising the functions that they might be familiar with" (MoNE, 2018a, p.

8), it can be claimed that the results may not support the increase in lexical complexity of the texts for grade 9. However, this result should be further examined because although grade 9 had significantly less diverse texts with medium to large effect size than above grades for CTTR, there were no significant differences between grade 9 and 12, and the significant differences between grade 9 and grades 10-11 had close-to-medium effect for MTLD. Similarly, grades 7 and 8 had lexically less diverse texts than grade 12, but the effect size was medium for each comparison. The results may be related to the sample size. Overall, the results for lexical diversity measures may indicate that, although there were some exceptions, grade levels 5, 7-8-9 and 10-11-12 tended to have texts with similar levels of lexical diversity.

As for lexical sophistication, only $CELEX_{logcontent}$ measure showed a significant difference between grades 7-8, 7-10-11-12 and 9-10 (with a medium effect size). No significant differences were found between the texts for grade 8 and the grades above it, between grade 9 and grades 11-12, and the cluster of 10-11-12. Similarly, $CELEX_{logall}$ did not show any differences with grade 7 and any grades above it, and only few differences between the grades above 7. The measure of LS-II did not show any significant differences between the grades, which was in line with the results of the previous studies (Lu, 2012). Similarly, the nonsignificant post hoc results for grade 5 and the above grades for lexical density measure (i.e., LD) aligned with the results of Lu (2012) in that it did not align with the scores of the raters for oral narratives of learners with different scores in that study as well.

Overall, when the results of linguistic complexity measures are considered, the results may show that the adjacent grades and a few clusters of grades tended to have texts with similar syntactic and lexical complexity levels although there were

differences between some grades that can be considered to function as ‘transition’ grades (e.g., grade 5 to 7 and grade 9 to 10) in some measures.

When the amount/hours of instruction corresponding to the clusters of grade levels are considered, the students are exposed to textbooks/input for CEFR level A1 for approximately 444 hours of 40-minute lessons, which consists of 222 hours for the grades 5-6, and 444 hours for the clusters of grades 7-8-9. Similarly, the amount of instruction for the grades 10-11-12 (B1 - B2+) are 444 hours of 40-minute lessons. In other words, the clusters of levels correspond to a period of 444 hours of exposure to the language/material in class. When the results of the complexity of the texts are interpreted considering the amount of exposure to the materials/language, it can be claimed that the students are exposed to texts/input with lower levels of complexity (CEFR A1-A2) for a longer period of time than the texts with higher levels of complexity (CEFR B levels).

5.1.2 Linguistic complexity of the listening texts

The second research question investigated whether there were significant differences among the listening texts in the textbooks for different grades in terms of their linguistic complexity. The results showed partial similarities with the results of the linguistic complexity measures of the reading texts across grades.

When the syntactic complexity of the texts is considered, it can be claimed that although grades 2,3 and 4 are considered CEFR A1 level along with grades 5 and 6, the syntactic complexity of the texts based on length-based measures (i.e., MLS, MLT, MLC) were significantly lower than all the grades above, which might be an indication of a slight increase in possible difficulty of the texts for the learners. However, no significant differences were found between grade 4 and grade 5 in

terms of their lexical diversity, so although there were longer sentences for listening texts starting from grade 5, the texts for grade 5 were not necessarily more lexically diverse than grade 4. However, starting from grade 6, the listening texts got lexically more diverse and had longer sentences/clauses than grade 4, which may be an indication of increase in the overall linguistic complexity.

Although there was a significant correlation in the nonparametric tests between subordination, coordination and the amount of complex noun phrases in the listening texts and the grade levels, the length-based measures showed mixed results. However, the results of the lexical complexity measures, more specifically lexical diversity measures, showed that the clusters of grades tended to have texts with similar lexical complexity levels as in the case of grades 4-5, 6-7-8, 7-8-9, 9-10-12, which was similar to the results of the lexical complexity of the reading texts to a degree. However, it was seen that grade 11 had more lexically diverse texts than all the other grades, so the listening texts might be more difficult for the learners than grade 12, which corresponded to a higher CEFR level. On a similar vein, the listening texts for grade 9 (CEFR A1/A2) was not significantly less complex than grade 10 (A2+/B1) and 12 (B2+) for measures of CTTR, SVV1, which is contrary to expectations.

5.1.3 Comparison of the textbook series

Although there have been studies aiming to set benchmarks for CEFR levels (Jin et al., 2020), there are still no specific thresholds or benchmarks for identifying the proficiency levels and grading the textbooks. Therefore, the texts in the present study were analyzed in terms of the degree of significant differences between grade and/or CEFR levels they were designed for. However, in order to have a better

understanding of the levels of the texts, textbooks for two grades (i.e., 11 and 12) were further compared to two textbooks of an internationally published textbook series with similar CEFR levels (i.e., Interchange3 and Passages1).

The results of the readability of the texts showed that although Coh-Metrix L2 Readability Index did not show any differences between the locally and internationally published textbooks, the traditional readability formulas showed that the textbook for grade 11 had easier/more readable texts than the internationally published textbooks, which was in line with the results of the linguistic complexity measures. Similarly, the textbook for grade 12 had texts that were similar to the ones in the two comparable textbooks in terms of their levels of readability.

Considering the results of the linguistic complexity measures for the texts in the textbooks, it can be claimed that the textbook for grade 11 had texts that were syntactically less complex than internationally published textbooks except for the measures of noun phrase complexity, left embeddedness and mean length of clauses, for which either no significant differences or differences with small effect sizes were found. However, the lexical complexity (i.e., diversity and sophistication) of the texts were similar to the internationally published textbooks of similar CEFR levels. It can be argued that the texts in the textbooks for grade 11 may need more syntactically complex sentences especially regarding sentence length, subordination and the number of dissimilar sentences across the texts. The textbook for grade 12, on the other hand, had texts that were similar to internationally published textbooks regarding their syntactic complexity except for the low amount of subordination and dissimilarity across sentences in the textbook for grade 12. When the medium effect sizes and the nonsignificant results for lexical diversity measures are considered, it can be argued that grade 12 had texts that were similar to the texts in Interchange3

and Passages1 (except for the significant result with medium effect size for CTTR). However, the texts had higher incidence of less frequent words than Interchange3, but not Passages1. Since the textbook for Grade 12 is expected to have texts with CEFR B2+ level, it can be argued that the texts for grade 12 may be linguistically less demanding than they are expected to be, so they may not be providing the input aimed for the level.

As for the listening texts, the textbook for grade 11 had listening texts that were similar to the internationally published textbooks in terms of their syntactic complexity, except that the texts for grade 11 had longer clauses and more complex nominals than Interchange3. The listening texts for grade 11 were also more lexically diverse than the texts in Interchange3 and more lexically sophisticated than both textbooks. Considering the mixed results for the measure of LS-II for the textbooks for state schools, it can be claimed that the listening texts in the textbook for grade 11 were, to a degree, more linguistically complex than a textbook corresponding to CEFR B1+, but similar to a textbook of CEFR B2. Since the textbooks for grade 11 are expected to be within the range of CEFR B1+/B2, it can be said that the textbook may provide linguistic input that would be within the range of CEFR level aimed. As for the texts for grade 12, the texts were similar to both textbooks in terms of their syntactic complexity. For lexical complexity, the texts were not more lexically diverse than the texts in Interchange3, but they had less variety of words than Passages1. Considering that the textbook for grade 12 is expected to have texts for CEFR B2+ level of proficiency, it can be argued that the listening texts may not be syntactically and lexically complex enough for the aimed level.

Regarding the comparisons of the textbooks, it should be noted that the results are not conclusive in that the internationally published textbooks are not the

sole or definite criterion for the textbooks to be compared to. However, the results of the comparisons among the textbooks aligned with the results of the comparisons among the grade levels in the state textbooks. It can be argued that the textbooks for grades 11 and 12 may not have reading texts that differ in their linguistic complexity and the listening texts for grade 11 are linguistically more complex than the texts for Grade 12.

5.2 Conclusion

The present study investigated the complexity of the texts in the L2 English textbooks used in state schools in Turkey from grades 2 to 12 through readability formulas and measures of linguistic complexity. The results showed that the complexity of the texts had both a linear and nonlinear increase in the syntactic and lexical complexity of the texts across grade levels, which was a finding partly similar to the findings by Jin et al. (2020). Both the reading and listening texts tended to remain stable across two or three adjacent grade levels (e.g., 2-3-4, (4)-5-6, 7-8-9, 10-11-12), despite some exceptions, for many measures of linguistic complexity. This finding was in line with the CEFR levels of the textbooks for primary and middle school in that the textbooks were expected to have CEFR A1 level texts for grade levels 2,3,4,5 and 6, A2 for 7,8 for middle school and grade 9 for high school as a revision. However, the linguistic complexity of the texts in the textbooks for high school did not show a gradual increase from Grade 10 to 12 although the textbooks for each grade level was assigned a different CEFR level ranging between A2/B1 to B2+.

The study further compared two locally published textbooks to two internationally published comparable textbooks. The results showed that the reading

texts for grade 11 may be less syntactically complex than they are expected to be for certain indices while the lexical complexity of the texts were mainly similar to the texts of similar CEFR levels, with some exceptions. The listening texts for grade 11 were similar to those of the internationally published textbooks, which may mean they were within the range of CEFR level the textbook was aimed for. As for grade 12, both the listening and reading texts were expected to be different from the internationally published textbooks, especially from Interchange3. With some exceptional results for certain measures, the textbooks mainly had texts with similar levels of linguistic complexity, which may indicate that the texts for grade 12 might be easier than expected.

Although the study did not compare the complexity of the listening and reading texts, it was seen that the level of and increase in the complexity of the reading texts for different grade levels did not always align with the level of complexity of the listening texts across similar grades as grade 11 had more linguistically complex listening texts than grade 10 and 12, but there were no differences between the grades in terms of reading texts the textbooks had.

5.3 Implications of the study

The study may offer several implications for material design and levelling of L2 English textbooks. The results of the study showed that the textbooks for primary and secondary schools had texts that mainly increased in their linguistic complexity along with the increase in the CEFR levels although there were some exceptions. The results and some exceptions may call for the consideration of several aspects in material design. Although the reading texts for grade 5 were lexically less diverse (i.e., CTTR) than the texts for grade 7, the length-based measures of syntactic

complexity did not show significant differences between grade 5 and 7 while grade 6 had texts with shorter sentences/T-units than grade 7 (an effect size close to large). In addition, the texts for grade 6 were shorter than grade 5. The longer texts (and the nonsignificant results) for grade 5 might have resulted from the inclusion of texts in the corpus because of the lack of clarity in the objectives for the level. These might indicate the need for more performance-based objectives for the primary school textbooks and a further consideration of the text length in the textbooks across grades.

Contrary to the results of the complexity of the texts for the primary and secondary schools, the textbooks for high school, especially for grades 10-12, did not increase in their readability and linguistic complexity, which may indicate that they may not be providing the input corresponding to the CEFR levels aimed. This highlights the need for attending to different subconstructs of linguistic complexity while designing the L2 materials for higher levels. In addition, although the present study did not focus on the variances in the genres used, the textbook for grade 11 tended to have a number of narrative texts, which might have affected the readability of the texts and resulted in more readable texts. Therefore, it can be claimed that the variety of texts with different genres used in the textbooks and their prevalence might need to be considered in the material design.

The results of the study also highlighted the need for the benchmarks for proficiency levels in that the finding that the listening texts for grade 11 were more complex than all the other levels may imply that the texts for grade 11 were above the CEFR level the grade was aimed for. However, comparing the texts to the internationally published textbooks showed that the texts may be within the range of

CEFR level, not necessarily more difficult, rather the listening texts for grade 12 might be easier for the target learners and the level.

The present study might further have some implications for the inclusion of measures to analyze linguistic complexity. In line with the previous studies (Jin et al., 2020), overall length-based measures aligned with the CEFR levels corresponding to the grade levels more than other measures of syntactic complexity. The sentence-syntax similarity measures of Coh-Metrix (SYNSTRUTa, SYNSTRUTt) also aligned with these results to a degree. In addition, sentence coordination measure (i.e., T/S) did not discriminate between the grade levels, which was both in line with (Lu, 2011) and contradicted (Jin et al., 2020) the previous research. As for the lexical complexity measures, apart from some mixed results for certain measures, the measures of lexical sophistication (i.e., LS-II) and lexical density (i.e., LD) could not discriminate between the grade levels, which supported the results of the previous studies (Lu, 2012).

The results of the study also highlighted the importance of having a multidimensional approach to linguistic complexity (Bulté and Housen, 2012; Ortega, 2003) and the inclusion of lexical complexity (Vajjala & Meurers, 2012) in the analysis considering the results that although some texts in the textbooks for adjacent levels may not differ in some syntactic complexity measures (e.g., listening texts for grades 5 and 6), there might be differences in some lexical complexity measures for the same texts (e.g., CTTR).

5.4 Limitations

There were several limitations to the study. One of the limitations of the study was about the tools used. Because of the interface and adaptability of the tools with the

software of the computers, the online versions of the tools were used, which required the texts to be analyzed separately and might have resulted in errors that might affect the reliability of the data. However, this was alleviated to a degree in that all the results were checked after each coding to the spreadsheet and some of the extreme outliers were checked and corrected in case of mistakes, which were very few in numbers.

5.5 Suggestions for future research

The results and the limitations of the study highlighted several areas that can be further investigated in future research regarding the complexity of the texts and levelling of the textbooks. As the study only focused on the texts in the textbooks, future research may also include the teachers who have used the textbooks through follow-up questionnaires or surveys. This would provide insights into the actual use of the textbooks and the perceptions of the teachers regarding the complexity of the textbooks. In addition, other aspects of text difficulty such as the learners' perceptions and the content of the texts can also be analyzed for the future studies.

The size of the sample/corpus might also be increased in future research. Several studies have investigated the textbooks used over a period of time (Jin et al., 2020), so the L2 English textbooks that have been used in Turkey for a certain period of time can be analyzed. Similarly, the textbooks used within specific periods of time can further be compared to other periods to see how the difficulty/complexity of the L2 English textbooks have changed over time in Turkey along with the changes in the curriculum towards a more communicative approach.

Future research can also focus on the differences in the linguistic complexity and readability of the texts based on different genres. As the genres used might have

an effect on the complexity of the texts (Lu, 2011; Polio & Yoon, 2018), analyzing text complexity in relation to the genres might provide further practical information on the design of the textbooks.

APPENDIX A

THE TEXTBOOKS IN THE CORPUS

Locally Published Textbooks

Grade / CEFR Level	Textbook	Publisher
2 / A1	İngilizce 2.Sınıf Ders Kitabı	Bilim ve Kültür Yayınları
3 / A1	Just Fun English	Tutku Yayıncılık
4 / A1	İngilizce Ders Kitabı 4	FCM Yayıncılık
5 / A1	Happy English	Başaran Yayıncılık
6 / A1	English Route	Monopol Yayıncılık
7 / A2	İngilizce Ders Kitabı 7. Sınıf	Bilim ve Kültür Yayınları
8 / A2	Upswing English	Tutku Yayıncılık
9 / A1-A2	Teenwise	T.C. Milli Eğitim Bakanlığı
10 / A2+-B1	Count Me In	T.C. Milli Eğitim Bakanlığı
11 / B1+-B2	Silver Lining	T.C. Milli Eğitim Bakanlığı
12 / B2+	Count Me In	T.C. Milli Eğitim Bakanlığı
Internationally Published Textbooks		
B1/B1+	Interchange3	Cambridge University Press
B2	Passages1	Cambridge University Press

APPENDIX B

EFFECT SIZES (Cohen's *d*) FOR

POST-HOC COMPARISONS AMONG GRADE LEVELS

Table B1. Effect Size (Cohen's *d*) for Post Hoc Comparisons for Readability Formulas among Grade Levels

Grade		Flesch Ease	Flesch Kincaid Grade Level	Coh-Metrix L2 Readability Index
9	10	1.22	1.42	1.31
	11	.49	.82	1.04
	12	1.36	1.60	1.66
10	11	-.68	-.63	-.19
	12	.38	.44	.39
11	12	.95	.98	.56

Table B2. Effect size (Cohen's *d*) for Post Hoc Comparisons for Measures of L2SCA among Grade Levels for Reading Texts

Grad		MLS	MLT	MLC	T/S	CN/C	DC/T
5	6	-.49	-.23	-.22	-.69	.25	-.38
	7	.58	.77	.48	-.51	1.11	.47
	8	1.45	1.47	1.25	-.18	1.55	.54
	9	1.42	1.62	.88	-.29	1.75	.62
	10	2.88	2.95	2.09	-.07	2.85	.77
	11	3.29	3.36	1.48	-.18	2.66	.76
	12	2.39	2.43	1.36	-.27	1.98	.82
6	7	.97	.98	.73	.27	.87	.60
	8	1.82	1.65	1.46	.77	1.37	.65
	9	1.86	1.87	1.19	.70	1.45	.69
	10	3.23	3.15	2.46	1.05	2.52	.81
	11	3.71	3.62	1.84	1.06	2.26	.76
	12	2.65	2.57	1.59	.77	1.78	.83
7	8	.77	.73	.85	.48	.66	.03
	9	.64	.62	.33	.38	.43	.26
	10	2.05	2.02	1.55	.72	1.04	.46
	11	2.17	2.17	.92	.65	.98	.74
	12	1.85	1.85	.95	.43	.97	.70
8	9	-.21	-.25	-.67	-.15	-.35	.21
	10	1.26	1.10	.30	.17	.43	.42
	11	1.24	1.09	-.22	.03	.01	.69
	12	1.25	1.20	.06	0.12	.22	.67
9	10	1.61	1.63	1.45	.37	1.03	.27

	11	1.68	1.77	.70	.24	.54	.65
	12	1.48	1.53	.77	.04	.64	.60
10	11	-.22	-.15	-.80	-.18	-.60	.31
	12	.26	.37	-.23	-.34	-.19	.30
11	12	.45	.51	.32	-.20	.27	.005

Note. r values are reported for the effect size for the comparisons for DC/T; MLS = Mean length of sentence; MLT = Mean length of T-unit; MLC = Mean length of clause; T/S = T-unit per sentence; CN/C = Complex nominal per clause; DC/T = Dependent clause per T-unit

Table B3. Effect size (Cohen's *d*) for Post Hoc Comparisons for Syntactic Complexity Measures of Coh-Metrix for Reading Texts among Grade Levels

Grade		READSL	SYNLE	SYNNP	SYNSTRUTa	SYNSTURTt
7	8	.89	.90	.93	.09	.39
	9	.74	.70	.74	-.21	-.02
	10	2.15	2.09	1.59	1.25	1.38
	11	2.08	1.91	1.54	1.08	1.37
	12	1.78	1.74	.99	1.18	1.42
8	9	-.25	-.40	-.13	-.30	-.48
	10	1.14	.99	.36	1.15	1.26
	11	.93	.90	.41	.97	1.25
	12	1.07	1.05	.08	1.08	1.31
9	10	1.53	1.62	.51	1.53	1.68
	11	1.37	1.46	.55	1.34	1.67
	12	1.32	1.42	.22	1.43	1.71
10	11	.34	-.05	.07	-.20	-.03
	12	-.17	.33	-.25	.06	.11
11	12	.45	.37	-.30	.22	.15

Note. READSL = Sentence length; SYNLE = Number of words before the main verb; SYNNP = Number of modifiers per noun phrase; STRUTa = Sentence syntax similarity for adjacent sentences; STRUTt = Sentence syntax similarity across paragraphs

Table B4. Effect size (Cohen's *d*) for Post Hoc Comparisons for Lexical Complexity Measures of Coh-Metrix for Reading Texts among Grade Levels

Grade		SVV1	CTTR	MTLD	$CELEX_{logcontent}$
5	7	.84	1.45	-	-
	8	1.14	2.23	-	-
	9	1.45	2.43	-	-
	10	3.18	4.19	-	-
	11	3.07	3.89	-	-
	12	2.64	3.49	-	-
7	8	.20	.82	.07	1.07
	9	.69	1.20	.36	.59
	10	2.26	2.88	1.27	1.62
	11	2.22	2.56	1.23	1.11
	12	2.00	2.15	.83	1.17
8	9	.54	.47	.32	-.50
	10	2.15	2.08	1.30	.25
	11	2.11	1.75	1.24	-.07
	12	1.90	1.35	.81	.05
9	10	1.40	1.43	.68	.88
	11	1.39	1.12	.69	.47
	12	1.35	.75	.43	.57
10	11	.03	-.34	.05	-.38
	12	.18	-.74	-.15	-.20
11	12	.15	-.39	-.18	.13

Note. CTTR = Corrected TTR; SVV1 = Squared verb variation; MTLD = Measure of Textual Lexical Diversity; $CELEX_{logcontent}$ = CELEX Log frequency for content words

Table B5. Effect size (Cohen's *d*) for Post Hoc Comparisons of Linguistic Complexity Measures for Listening Texts among Grade Levels

Grade		MLS	MLT	MLC	CTTR	SVV1
2	3	.75	.54	.54	-	-
	4	.81	.54	.71	-	-
	5	2.09	1.93	1.56	-	-
	6	2.24	2.06	1.71	-	-
	7	2.92	2.60	2.20	-	-
	8	1.75	2.09	1.86	-	-
	9	2.60	2.74	2.76	-	-
	10	2.20	2.00	2.12	-	-
	11	4.45	4.37	4.54	-	-
	12	2.46	2.22	2.06	-	-
3	4	.01	-.07	.12	-	-
	5	1.26	1.09	.98	-	-
	6	1.38	1.41	1.03	-	-
	7	2.35	2.16	1.72	-	-
	8	1.29	1.63	1.24	-	-
	9	1.63	1.97	1.78	-	-
	10	1.85	1.75	1.67	-	-
	11	3.65	3.66	2.93	-	-
	12	2.10	1.96	1.64	-	-
4	5	1.29	1.35	.91	.90	.56
	6	1.42	1.62	.96	1.94	1.22
	7	2.39	2.31	1.67	2.86	2.13
	8	1.30	1.77	1.18	2.19	1.55
	9	1.69	2.27	1.73	3.98	2.29
	10	1.86	1.81	1.62	3.60	2.91
	11	3.74	3.99	2.97	6.68	4.58
	12	2.11	2.03	1.59	3.24	2.31
5	6	.10	.59	-.04	1.21	.85
	7	1.45	1.62	.86	2.12	1.64
	8	.49	1.02	.22	1.51	1.18
	9	.26	1.17	.46	3.43	1.99
	10	1.25	1.36	.86	3.04	2.57
	11	2.54	3.14	1.26	6.53	4.30
	12	1.46	1.58	.85	2.66	2.05
6	7	1.38	1.13	.95	.45	.41
	8	.43	.53	.29	.33	.27
	9	.14	.44	.58	1.98	1.13
	10	1.20	1.06	.94	1.71	1.53
	11	2.47	2.31	1.52	4.28	3.26
	12	1.41	1.27	.93	1.46	1.28
7	8	-.69	-.55	-.68	.00	-.09
	9	-1.32	-.85	-.60	2.04	.89
	10	.23	.25	.02	1.67	1.32
	11	.63	.70	.07	5.70	3.21

	12	.37	.45	.04	1.34	1.07
8	9	-.35	-.21	.20	1.54	.88
	10	.76	.67	.68	1.31	1.26
	11	1.33	1.39	.98	3.58	3.02
	12	.91	.87	.68	1.09	1.06
9	10	1.15	.86	.60	-.16	.28
	11	2.45	2.01	1.06	1.97	2.01
	12	1.36	1.08	.60	-.31	.22
10	11	.20	.23	.10	2.00	1.84
	12	.10	.16	.01	-.15	-.02
11	12	-.07	-.01	.11	-2.01	-1.67

Note. MLS = Mean length of sentence; MLT = Mean length of T-unit; MLC = Mean length of clause; CTTR = Corrected TTR; SVV1 = Squared verb variation

Table B6. Effect size (Cohen's *d*) for Post Hoc Comparisons of Syntactic Complexity of Measures of L2SCA for Reading Texts for Comparison

Grade		MLS	MLT	DC/T	T/S
11	Int3	-1.39	-.81	-1.16	-.92
	Pass1	-.82	-.82	-.88	-.20
12	Int3	-.49	-.12	-1.16	-1.01
	Pass1	-.13	.00	-.91	-.37

Note. MLS = Mean length of sentence; MLT = Mean length of T-unit; MLC = Mean length of clause; DC/T = Dependent clause per T-unit; T/S = T-unit per sentence

Table B7. Effect size (Cohen's *d*) for Post Hoc Comparisons of Syntactic Complexity Measures of Coh-Metrix for Reading Texts for Comparison

Grade		READSL	SYNLE	SYNNP	SYNSTRUTa	SYNSTRUTt
11	Int3	-1.58	-.64	.61	1.26	1.24
	Pass1	-.88	-.27	.67	.76	.87
12	Int3	-.61	-.13	.58	.80	.94
	Pass1	-.15	.19	.25	.06	.63

Note. Int3 = Interchange3; Pass1 = Passages1; READSL = Sentence length; SYNLE = Number of words before the main verb; SYNNP = Number of modifiers per noun phrase; SYNSTRUTa = Sentence syntax similarity for adjacent sentences; SYNSTRUTt = Sentence syntax similarity across paragraphs

Table B8. Effect size (Cohen's *d*) for Post Hoc Comparisons of Lexical Complexity Measures for Reading Texts for Comparison

Grade		CTTR	$CELEX_{logall}$	LS-II	LD	LV
11	Int3	.29	-.67	.94	-.38	.99
	Pass1	-.36	.52	.37	-.79	.63
12	Int3	-.21	-1.03	.77	.10	.47
	Pass1	-.84	.08	.13	-.22	.10

Note. Int3 = Interchange3; Pass1 = Passages1; CTTR = Corrected TTR; MTLD = Measure of Textual Lexical Diversity; $CELEX_{logcontent}$ = CELEX Log frequency for content words; LS-II = Lexical sophistication; LD = Lexical density; LV = Lexical variety

Table B9. Effect size (Cohen's *d*) for Post Hoc Comparisons of Lexical Complexity Measures for Listening Texts for Comparison

Grade		LS-II	CTTR	SVV1	LV
11	Int3	2.75	2.31	1.35	.50
	Pass1	1.94	.64	.46	1.99
12	Int3	1.78	-.28	-.49	.84
	Pass1	1.11	-1.15	-1.03	1.94

Note. Int3 = Interchange3; Pass1 = Passages1; LS-II = Lexical sophistication; CTTR = Corrected TTR; SVV1 = Squared verb variation; LV = Lexical variety

REFERENCES

Ağçam, R. & Babanoğlu, M. P. (2018). A comparative study on EFL textbooks in Turkish and German secondary public schools. *International Journal of Eurasia Social Sciences, 9*(32), 948-959.

Anthony, L. (2021). AntFileConverter (Version 2.0.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Arıkan, A. (2007). Pragmatic problems in elementary level ELT coursebooks: Focus on dialogues. In E. Weigand (Ed.), *Dialogue Analysis XI: Proceedings of the 11th IADA Conference on 'Dialogue Analysis and Rhetoric'* (pp. 3-13), Münster, Germany: University of Münster.

Aslan, H. İ., Keskin, S. C., & Önder, S. (2019). Values in the 6th grade English textbooks. *Journal of the International Scientific Research (IBAD), 4*(2), 355-371. DOI: 10.21733/ibad.551643

Aşık, A. (2017). A sample corpus integration in language teacher education through coursebook evaluation. *Journal of Language and Linguistic Studies, 13*(2), 728-740.

Ayaz, A. D., Ozkardas, S., & Ozturan, T. (2019). Challenges of English language teaching in high schools in Turkey and possible suggestions to overcome them. *Eurasian Journal of Applied Linguistics, 5*(1), 41–55. Doi:10.32601/ejal.543778

Bailin A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: a critique. *Language & Communication 21*, 285–301.

Batur Z., & Özcan, H. (2020) Readability levels of texts in the sixth grade Turkish textbooks. *Azerbaijan Journal of Educational Studies, 690*(1), 217-230. http://dx.doi.org/10.29228/edu.103

Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools, 51*(2), 198-215. DOI: 10.1002/pits.21740

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development.* Boston, Massachusetts: Heinle & Heinle Publishers.

Brown, J. D. (1998). An EFL Readability Index. *JALT Journal, 20*(2), 7-36.

Bui, G., & Skehan, P. (2018). Complexity, accuracy and fluency. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). John Wiley & Sons, Inc. DOI: 10.1002/9781118784235.eelt0046

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam, The Netherlands: Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26,* 42–65. http://dx.doi.org/10.1016/j.jslw.2014.09.005

Bush, L., Koons, H., & Sanford-Moore, E. E. (2016). *Primary and secondary English textbook complexity in the Republic of Korea.* Retrieved from https://metametricsinc.com/research-publications/examination-text-complexity-efl-graded-readers/?full_article=true

Byrd, P. (2001). Textbooks: Evaluation for selection and analysis for implementation. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed., pp. 415-428). Boston: Heinle & Heinle/Thompson Learning.

Carroll, J. B. (1964). *Language and thought.* Englewood Cliffs, NJ: Prentice-Hall.

Chen, A. C.-H. (2016). A critical evaluation of text difficulty development in ELT textbook series: A corpus-based approach using variability neighbor clustering. *System, 58*, 64-81. https://doi.org/10.1016/j.system.2016.03.011

Crawford, J. (1995). The role of materials in the language classroom: Finding the balance. *TESOL in Context, 5*(1), 25-33.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475–493.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84-101.

Crossley, S. A., & McNamara, D. S. (2012). Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In S. Jarvis, & S. A. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach,* (pp. 106-126). Bristol, Blue Ridge Summit: Multilingual Matters.

Çakır, İ. (2010). The frequency of culture-specific elements in the ELT coursebooks at elementary schools in Turkey. *Novitas-ROYAL (Research on Youth and Language), 4*(2), 182-189.

Çetinkaya, G., Yenmez, A. A., Çelik, T., & Özpinar, İ. (2018). Readability of texts in secondary school Mathematics course books. *Asian Journal of Education and Training, 4*(4), 250-256. https://doi.org/10.20448/journal.522.2018.44.250.256

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 27*(1), 11-20+28.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English, 26*(1), 19-26.

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17*(2), 187-209.

Dufty, D. F., Graesser, A. C., Louwerse, M. M., & McNamara, D. S. (2006). Assigning grade levels to textbooks: Is it just readability? *Proceedings of the Annual Meeting of the Cognitive Science Society, 28*(28), 1251-1256.

Ellis., R. (2005). Principles of instructed language learning. *System, 33,* 209-224. doi:10.1016/j.system.2004.12.006

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Field, A. (2009). *Discovering statistics using SPSS* (3rd Ed.). London: Sage Publications.

Flesch, R. (1948). A new readability yardstick*. *Journal of Applied Psychology, 32*(3), 221-233.

Fry, E. (1969). A readability formula that saves time. *Journal of Reading, 11*(7), 513-516, 575-578.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher, 56*(3), 286-291.

Gedik, T. A. (2020). Motion lexicon: A corpus-based comparison of English textbooks and university entrance exams in Turkey. *Journal of Foreign Language Teaching and Translation Studies, 5*(3), 31-44. DOI: 10.22034/efl.2020.246122.1053

Green, C. (2019). A multilevel description of textbook linguistic complexity across disciplines: Leveraging NLP to support disciplinary literacy. *Linguistics and Education, 53,* 1-11. https://doi.org/10.1016/j.linged.2019.100748

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal, 26*(1), 5-24.

Guiraud, P. (1960). *Probl`emes et me´thodes de la statis- tique linguistique [Problems and methods of sta- tistical linguistics].* Dordrecht, The Netherlands: D. Reidel.

Halliday, M. A. K. (2008). *Complementarities in language.* Beijing: The Commercial Press.

Harwood, N. (2016). What can we learn from mainstream education textbook research? *RELC Journal, 48*(2), 264-277. https://doi.org/10.1177/0033688216645472

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA* (pp. 1-20). Amsterdam, The Netherlands: Benjamins.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Boston, MA: Pearson.

Hughes, S. H. (2019). Coursebooks: Is there more than meets the eye*? ELT Journal, 73*(4), 447-455. https://doi.org/10.1093/elt/ccz040

Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the output hypothesis. *Applied Linguistics, 24*(2), 168-196.

Jin, T., Lu, X., & Ni, J. (2020). Syntactic complexity in adapted teaching materials: Differences among grade levels and implications for benchmarking. *The Modern Language Journal, 104*(1), 192-208. DOI: 10.1111/modl.12622

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing, 37*(13), 13-38. DOI:10.1016/j.jslw.2017.06.001

Jordan, G., & Gray, H. (2019). We need to talk about coursebooks. *ELT Journal, 73*(4), 438-446. https://doi.org/10.1093/elt/ccz038

Kalantari, R., & Gholami, J. (2017). Lexical complexity development from dynamic systems theory perspective: Lexical density, diversity, and sophistication. *International Journal of Instruction, 10*(4), 1-18. https://doi.org/10.12973/iji.2017.1041a

Kalayci, D. (2018). *Yabanci dil olarak Türkçe ve Ingilizce öğretimi ders kitaplarinin karşilaştirilmasi* (Unpublished MA Thesis). Karadeniz Teknik Universitesi.

Karakus, B., Aydın, G., & Hallac, I. R. (2015). *Distributed readability analysis of Turkish elementary school textbooks.* Proceedings of International Conference on Information Technology and Computer Science. Retrieved from https://arxiv.org/pdf/1802.03821.pdf

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*(1), 1-22. DOI:10.1111/modl.12447

Kirkgöz, Y. (2007). Language planning and implementation in Turkish primary schools. *Current Issues in Language Planning, 8*(2), 174-191. DOI:10.2167/cilp114.0

Kizildag, A. (2009). Teaching English in Turkey: Dialogues with teachers about the challenges in public primary schools. *International Electronic Journal of Elementary Education, 1*(3), 188-201.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10*(1), 62-102.

Koizumi, R., & In'ami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40,* 554-564. https://doi.org/10.1016/j.system.2012.10.012

Koslin, B. I., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading.* New York: College Entrance Examination Board.

Köroğlu, Z. Ç., & Elban, M. (2020). National and global identity perspectives of textbooks: Towards a sense of global identity. *Advances in Language and Literary Studies, 11*(5), 55-65. http://dx.doi.org/10.7575/aiac.alls.v.11n.5-p.55

Krashen, S. (1985). *The input hypothesis: Issues and implications.* London: Longman.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York, NY: Routledge.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*(S1), 127-159. https://doi.org/10.1111/lang.12115

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

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*, 21–33.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical Richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Littlejohn, A. (2011). The analysis of language teaching materials: Inside the Trojan Horse. In B. Tomlinson (Ed.), *Materials development in language teaching* (2nd ed., pp. 179-211). Cambridge: Cambridge University Press.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). New York, NY: Academic Press.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496. doi 10.1075/ijcl.15.4.02lu

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly, 45*(1), 36-62.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal, 96*(2), 190-208. DOI: 10.1111/j.1540-4781.2011.01232.x

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing, 34*(4), 493-511. https://doi.org/10.1177/0265532217710675

Lu, X., Gamson, D. A., & Eckert, S. A. (2014). Lexical difficulty and diversity of American elementary school reading textbooks: Changes over the past century. *International Journal of Corpus Linguistics, 19*(1), 94-117. doi 10.1075/ijcl.19.1.04lu

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment.* New York, NY: Palgrave Macmillan.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the Measure of Textual, Lexical Diversity* (Unpublished PhD Thesis). The University of Memphis.

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

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading, 12*(8), 639-646.

McLaughlin, B. (1987). *Theories of second language learning.* London: Edward Arnold.

McNamara, D. S., Graesser, A., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In J.P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Lanham, MD: R&L Education.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* New York, NY: Cambridge University Press.

Mert, E. L. (2013). The readability of the texts in the Turkish textbooks in Turkey. *Mersin University Journal of the Faculty of Education, 9*(3), 87-98. DOI:10.17860/efd.14421

Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics, 32*, 111–129. doi: 10.1017/S0267190512000128

Ministry of National Education (2012). *Millî eğitim bakanliği ders kitaplari ve eğitim araçlari yönetmeliği.* Retrieved from http://mevzuat.meb.gov.tr/dosyalar/1605.pdf

Ministry of National Education. (2018a). *İngilizce dersi öğretim programı (ilkokul ve ortaokul 2, 3, 4, 5, 6, 7 ve 8. Sınıflar)*. Ankara, Turkey: Author.

Ministry of National Education. (2018b). *İngilizce dersi öğretim program (9,10, 11 ve 12. Sınıflar)*. Ankara, Turkey: Author.

Ministry of National Education. (2019). *Mutlu Çocuklar, Güçlü Türkiye: 2023 EğitimVizyonu.* Ankara, Türkiye: Author.

Morales, B. C. (2019). Readability and types of questions in Chilean EFL high school textbooks. *TESOL Journal, 11*(2), 1-15. DOI: 10.1002/tesj.498

Mulyanti, W., & Soeharto, P. P. (2020). *Text complexity in English textbooks for junior high school: A systemic functional perspective.* Paper presented at Twelfth Conference on Applied Linguistics (CONAPLIN 2019).

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/ La Revue canadienne des langues vivantes, 63*(1), 59-81. DOI:10.1353/cml.2006.0049

Nation, I. S. P., & Heatley, A. (1994). Range: A program for the analysis of vocabulary in texts [software]. Retrieved from https://people.wgtn.ac.nz/paul.nation

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. doi:10.1093/applin/amp044

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

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/10.1111/lang.12079

Polio, C., & Yoon, H-J. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics, 28*(1), 165-188. https://doi.org/10.1111/ijal.12200

Putra, D. A., & Lukmana, I. (2017) Text complexity in senior high school English textbooks: A systematic functional perspective. *Indonesian Journal of Applied Linguistics, 7*(2), 436-444. doi: dx.doi.org/10.17509/ijal.v7i2.8352

Radić-Bojanić, B. B., & Topalov, J. P. (2016). Textbooks in the EFL classroom: Defining, assessing and analyzing. *Collection of Papers of the Faculty of Philosophy, 46*(3), 137-153. DOI:10.5937/ZRFFP46-12094

Richards, J. C. (n.d.). The role of textbooks in a language program. Retrieved from https://www.professorjackrichards.com/articles/

Richards, J. C., & Sandy, C. (2015). *Passages 1* (3rd ed.). New York, NY: Cambridge University Press.

Richards, J. C., Hull, J., & Proctor, S. (2017). *Interchange 3* (5th ed.). New York, NY: Cambridge University Press.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*, 26-43. DOI: 10.1111/j.1540-4781.2011.01146.x

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532. doi:10.1093/applin/amp047

Stenner, A. J. (1996). *Measuring reading comprehension with the Lexile Framework.* Paper presented at the North American Conference on adolescent/Adult Literacy (4th, Washington, DC., February 1996). Retrieved from https://files.eric.ed.gov/fulltext/ED435977.pdf

Sung, Y.-T., Dyson, S. B., Chen, Y-C., Lin, W-C., & Chang, K-E. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal, 99*(2), 371-391. DOI: 10.1111/modl.12213

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125- 144). Oxford: Oxford University Press.

Şener, S., & Mulcar, V. (2018). An investigation of teachers' perceptions on English textbooks: A case study of teachers teaching 10th graders in Mugla. *Western Anatolia Journal of Educational Sciences, 9*(1), 15-37.

Şimşek, M. R., & Dündar, E. (2016). *Exploring the pros and cons of a local English coursebook in user preferences.* Proceedings of INTCESS2016 3rd International Conference on Education and Social Sciences, 307-317. http://www.ocerint.org/intcess16_epublication/papers/177.pdf

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Tekir, S., & Arikan, A. (2007). An analysis of English language teaching coursebooks by Turkish writers: "Let's Speak English 7". *International Journal of Human Sciences, 4*(2), 1-18.

Thornbury, S. (2013). Resisting coursebooks. In J. Gray (Ed.), *Critical perspectives on language teaching materials* (pp. 204-224). New York, NY: Palgrave Macmillan.

To, V. (2017). Grammatical intricacy in EFL textbooks. *International Journal of English Language Education, 5*(2), 127-140. doi:10.5296/ijele.v5i2.12087

To, V. (2018). Linguistic complexity analysis: A case study of commonly-used textbooks in Vietnam. *SAGE Open, 8*(2), 1-13. https://doi.org/10.1177/2158244018787586

Tok, H. (2010). TEFL textbook evaluation: From teachers' perspectives. *Educational Research and Review, 5*(9), 508-517.

Tomlinson, B., & Masuhara, H. (2013). Adult coursebooks. *ELT Journal, 67*(2), 233-249. https://doi.org/10.1093/elt/cct007

Tüm, G., & Emre-Parmaksız, G. (2017). Comparison of speaking activities in Turkish and English language teaching coursebooks regarding self-assessment grid of CEFR. *Journal of Language and Linguistic Studies, 13*(2), 367-378.

Turkben, T. (2019). Readability characteristics of texts in middle school Turkish textbooks. *Educational Policy Analysis and Strategic Research, 14*(3), 80-105. doi: 10.29329/epasr.2019.208.5

Uğurlu, M., & Taş, S. (2020). The representation of cultures in English language textbooks: A comparison of three textbooks used in Turkey. *Ahi Evran Akademi, 1*(2), 54-67.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–173). Montréal, Canada: Association for Computational Linguistics.

Valizadeh, M. (2021). The challenges facing English language teachers in Turkey. *Advances in Language and Literary Studies, 12*(4), 61-67. http://dx.doi.org/10.7575/aiac.alls.v.12n.4.p.61

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998*). Second language development in writing: Measures of fluency, accuracy, and complexity.* Honolulu, HI: University of Hawai'i Press.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28,* 53-67. http://dx.doi.org/10.1016/j.jslw.2015.02.002

Yılar, M. B. (2020). 9. sınıf Coğrafya ders kitabında yer alan metinlerin okunabilirlik düzeyinin incelenmesi. *The Journal of Turkish Educational Sciences, 18*(2), 1126-1146. https://doi.org/10.37217/tebd.812110

Yüce, E., & Mirici, İ. H. (2009). A qualitative inquiry into the application of 9th grade EFL program in terms of the CEFR. *Journal of Language and Linguistic Studies, 15*(3), 1171-1187.

Zamanian, M., & Heydari, P. (2012). Readability of texts: State of the art. *Theory and Practice in Language Studies, 2*(1), 43-53. doi:10.4304/tpls.2.1.43-53

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47,* 1-15. https://doi.org/10.1016/j.asw.2020.100505