T.C

YEDİTEPE UNIVERSITY

GRADUATE INSTITUTE OF EDUCATIONAL SCIENCES

A CORPUS-BASED APPROACH TO TURKISH EFL TEXTBOOK EVALUATION: SINGLE WORD AND FOUR-WORD LEXICAL BUNDLE FREQUENCY

by

Kübra ÇİNAR

Submitted to the Graduate Institute of Educational Sciences
In partial fulfillment of the requirements for the degree of
Master Arts
in English Language Education

ISTANBUL, 2015

T.C.
YEDİTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

TEZ TESLİM ve ONAY TUTANAĞI

KONU: A Corpus-based Approach to Turkish EFL Textbook Evaluation: Single Word and Four-word Lexical Bundle Frequency

ONAY:

Yrd. Doç. Dr. Erkan Karabacak	
(Danışman)	(İmza)
Yrd. Doç. Dr. Zeynep B. Koçoğlu	
(Üye)	(İmza)
Yrd. Doç. Dr. Zeynep Canlıbel	
(Üye)	(İmza)

TESLİM EDEN : Kübra Çinar
TEZ SAVUNMA TARİHİ :20.../.04../2015
TEZ ONAY TARİHİ :20./.04./2015

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis supervisor Asst. Prof. Dr. Erkan Karabacak for his invaluable guidance, encouragement and patience throughout the whole process of writing this thesis. I appreciate his support both personally and academically with his in-depth expertise in corpus linguistics.

I would also like to thank to Asst. Prof. Dr. Zeynep B. Koçoğlu, my advisor in the M.A. program, for enlightening me with her advice and rich knowledge in English Language Teaching.

I owe special thanks to my father and my sister, who have always been supportive physically and emotionally. Without my father's time he spent for taking care of my beloved little daughter, and my sister's encouragement and guidance, it would not be possible to focus on my study.

Finally, my loving thanks go to my husband Şenol, who has shared life with me and given me love and more, and to my little angle Azra for providing the strength I need. I am so lucky to have you.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

EFL	English as a Foreign Language
MoNE	Turkish Ministry of National Education
TC-Tr	Textbook Corpus of ELT textbooks for Grades 4, 5, 6, 7 and 8, which are approved by the Turkish MoNE
TC-Ts	Textbook Corpus of *Touchstone 1*
OANC	Open American National Corpus
OANC-spoken	Spoken component of the OANC
NS	Native Speaker
BNC	British National Corpus
COCA	Corpus of Contemporary American English
CANCODE	Cambridge and Nottingham Corpus of Discourse in English
ESL	English as a Second Language
ELT	English Language Teaching
CEFR	Common European Framework of Reference for Languages

LIST OF TABLES

LIST OF FIGURES

ÖZET

Ders kitapları sınıf içi eğitimde önemli bir yer tutmaktadırlar ve bir değerlendirmeden geçirilmeleri de bu yüzden oldukça önemlidir. Türkiye'de çok büyük bir öğrenci kitlesi tarafından okunan ve birçoğu için temel kaynak olan Türk Milli Eğitim Bakanlığı tarafından onaylı, ilkokul ve ortaokul seviyesindeki (4., 5., 6., 7. ve 8. sınıflar) İngilizce Yabancı Dil öğretimi ders kitaplarının değerlendirilmesi elzemdir. Bu kitapların konuşma dili içeriğini değerlendirmek için, derlem tabanlı bir değerlendime yolu benimsendi. Bu nedenle de Türkiye'deki 11 İngilizce öğretim öğrenci kitabı ile bunların alıştırma ve öğretmen kitaplarında yer alan diyaloglardan bir derlem oluşturuldu. Ayrıca, aradaki farklılıkları çalışmak ve derlem-tabanlı ders kitaplarının gerçekten konuşma dilini daha iyi yansıtıp yansıtmadığını görmek için, derlem-tabanlı bir ders kitabı olan *Touchstone 1* kitabı derlemi oluşturuldu. Açık Amerikan Ulusal Derlemi'nin konuşma bölümündeki ilk üç binlik bantta yer alan sözcükleri ve dörtlü sözcük gruplarını referans listesi olarak kullanarak, bunların ders kitapları derlemlerindeki kullanım sıklıkları ve bunların her bir kitabın ne kadarını kapsadığı hesaplandı. Sonuçlar göstermiştir ki sınıf seviyesi yükseldikçe Türkiye'deki İngilizce ders kitaplarındaki sözcük seviyesinde düzenli bir artış , referans listesi sözcüklerini çok düşük seviyede içerdiğini gösterirken; *Touchstone 1*'in belirgin bir farklılık olmasa da, aynı seviyedeki Türkiyedeki kitaplara kıyasen daha fazla sözcük tipi içerdiğini ve bunları daha etkili bir şekilde tekrar ettiğini göstermektedir.

Anahtar sözcükler: Derlem-tabanlı ders kitabı, ders kitabı değerlendirme, otantiklik, ders kitabı İngilizcesi, konuşma İngilizcesi, yüksek sıklıktaki sözcükler, dörtlü sözcük grupları.

ABSTRACT

Textbooks play a key role in language classrooms, and their evaluation is of great importance. The evaluation of the English as a Foreign Language (EFL) textbooks approved by the Turkish Ministry of National Education and studied in elementary and secondary state schools (4th, 5th, 6th, 7th, and 8th grades) in Turkey is essential as they are in use on a large scale and the primary source for many students. In order to evaluate the spoken language content of those textbooks, a corpus-based approach as an evaluation method was adopted, and a corpus of dialogues in 11 Turkish EFL course books in addition to their workbooks and teacher books was created. Similarly, corpus of dialogues from *Touchstone 1*, a corpus-based textbook, was also compiled to study the differences between the two corpora regarding whether corpus-based textbooks better reflect authentic spoken language. Using single words and four-word bundles from the first three 1000 bands in the spoken part of the Open American National Corpus (OANC) as a reference list, frequencies and percentages of their use in the textbook corpora were calculated. The results showed that while the EFL textbooks in Turkey cover the words in reference lists at a very low level without a gradual development as the grade levels increase, the *Touchstone 1* covers more types and recycles them more effectively compared to the textbooks with corresponding grades, though the difference in formulaic language use was not outstanding.

Keywords**:** Corpus-based textbook, textbook evaluation, authenticity, textbook English, spoken language, high frequency vocabulary, four-word bundles.

I. INTRODUCTION

Over the past few decades, the interest in corpus studies has grown rapidly. Thanks to this growth, applied linguists and language practitioners can be more informed about the native speaker (NS) interactions, and have the chance of making use of them in their language practice. The developments in the computer technology have also rendered the compilation of corpora an easier task. Through the corpora software available to researchers, one can investigate the details of a language, especially English, which is studied by millions of people all over the world. Although it is a hot debate whether those learners should take NS norms as a base, it is still widely accepted by many. On the other hand, material designers are expected to serve for this preference.

While designing materials, including language learning textbooks, authors no longer need to rely on their intuition. Corpora provide them with what they lack, a well-developed body of knowledge (Dubin, 1995), so that they can see what hundreds of different speakers or writers have already said or written. Consequently, the materials utilizing corpora, such as corpus-informed textbooks, are expected to be more authentic, and give the sense of being realistic to learners.

On the other hand, in addition to designing materials by benefiting from corpus, it is also possible to evaluate the existing pedagogical descriptions in the light of corpus. Obviously, teachers and educational evaluators need to make choice among hundreds of language learning materials available in the market. While doing this, they have to compare and contrast textbooks, a prominent source of language learning, with each other considering some factors such as their grammar and vocabulary content as well as their organizations. In order to evaluate to what extent they are successful in including realistic input, NS corpora provide a great source while identifying what is and what is not real in the language since what has already been said or written is already there.

One empirical basis that could be used while making comparisons among textbooks and evaluating them is frequency information. Römer (2011) argues that confronting with the repeatedly occurring lexis and lexical combinations will help learners develop their receptive and productive skills. Considering this, textbook evaluators can take the language items that occur frequently in real-life situations into consideration in the evaluation process.

Schmitt and Schmitt (2012, p. 7) suggest that the knowledge of the most frequent 3000 word families is required at least "to largely understand (and presumably produce) conversational English". Consequently, the identification of frequently used words in real contexts is essential for language learners to know. As vocabulary knowledge is frequently associated with language proficiency, it is one of the most fundamental components in textbooks (Criado & Sánchez, 2009). Moreover, longer lexical sequences have also a significant role in language teaching (Cowie, 1992). Lexical bundles, defined as "the most frequently recurring lexical sequences" by Biber and Conrad (1999, p. 183), can also be referred to evaluate the representativeness of the authenticity of a language learning resource. Jablonkai (2009) argues that "three-word lexical bundles are often the ones which are part of four-word bundles, and four-word bundles are more frequent and give more variety for the structural and functional analysis than five-word bundles" (p. 4). In the current study, therefore, it is aimed to focus on four-word lexical bundles' frequency, in addition to single words' frequency.

Moreover, it seems that researchers have valued the use of corpus for pedagogical contexts by examining learner output in the light of a reference corpus (Nesselhauf, 2005; Shirato & Staplaton, 2007; Durrant & Schmitt, 2010). In non-native contexts like Turkey, EFL textbooks are the primary source of language for many learners. Therefore, textbooks represent the language input provided for learners in language classrooms to a large extent.

Ideally, English language teachers are supposed to teach learners authentic English so that they can communicate easily and effectively. Reppen (2012) contends that "if a feature is very common, and is used by fluent native speakers of English, then we should certainly teach that feature to learners" (p. 14). In order to understand to what extent students are exposed to common English at vocabulary level that they might be confronted with in real life, an EFL textbook corpus would be of great help. After compiling the corpus, thanks to available corpus software, it takes little time to evaluate whether textbooks can mirror authentic English in many aspects, including the frequency of occurrence of single lexical items and multi-word combinations.

Despite its recognised strengths in language teaching/learning context, it seems that pedagogical corpus applications have been neglected by teachers, learners and language authorities such as administrators and material writers. Corpus-based studies so far have revealed significant facts about English language classroom input. Among them, studies examining language textbooks as well as teacher talk indicate that there are discrepancies between the input and general corpora in several aspects. Cullen and Kuo (2007) report that English as a Foreign Language (EFL) textbooks published in the United Kingdom since the year 2000 cover features of spoken grammar inadequately. They argue that there is a lack of awareness of corpora in pedagogical practice. As a primary source for English language learners in Turkish context, textbooks widely used in Turkey are also worth of deep investigation and evaluation. The main motivation behind the present study is to reveal whether it might be true that compared to a NS corpus, EFL textbooks in Turkey do not adequately cover aspects of spoken English language. By surveying the EFL textbooks studied in elementary and secondary state schools (4th, 5th, 6th, 7th and 8th grades), which are in use on a large scale in Turkey, the current study basically investigates whether there is a gap in their lexical choices compared to the lexical items used in real conversational language.

A corpus analysis of textbooks contribute to the body of literature in English language teaching in Turkish context, as it indicates whether the authors' selection of lexical content represents an authentic use of English. Such an analysis also may be helpful to identify lexical content differences, if any, among those textbooks. Moreover, the same kind of an analysis may indicate how successful a corpus-informed textbook is in representing authentic English. Keeping these in mind, a corpus analysis of textbooks might also give an idea about the quality and quantity of the language used by Turkish learners. It seems that in Turkish context little research has been conducted to identify the possible gaps between NS corpora and learner input. Therefore, the present study aims to contribute to the literature in search of an answer to the following questions:

1. Does the English taught at Turkish state schools represent the English which is used by native speakers?

   - To what extent do the textbooks approved by the MoNE for 4th, 5th, 6th, 7th and 8th grades represent authentic spoken English regarding the most frequently used 1st, 2nd and 3rd 1000 word bands?

   - To what extent do they represent authentic spoken English regarding the most frequently used 1st, 2nd and 3rd 1000 four-word lexical bundle bands?

2. Are the results different for *Touchstone 1*, a corpus-informed textbook?

   - To what extent does *Touchstone 1* represent authentic spoken English regarding the most frequently used 1st, 2nd and 3rd 1000 word bands?

   - To what extent does it represent authentic spoken English regarding the most frequently used 1st, 2nd and 3rd 1000 four-word lexical bundle bands?

II. LITERATURE REVIEW

This chapter begins with the definition of *corpus methodology*. Then, highlighting its indirect use, the benefits of corpus methodology in language teaching and learning is discussed. In the light of the research studies based on the comparison of EFL textbook corpora with native speaker corpora, the relevance of corpora as a tool to evaluate textbooks available in the market is pointed out. As the studies indicate a gap between 'textbook' English and authentic English, the issue of authenticity in textbooks is also discussed. Finally, focusing on vocabulary and lexical bundles in spoken corpus, the importance of frequency is addressed.

2.1. Corpus and its Types

A corpus is simply stated as a large collection of texts (McCarthy, 2004). Leech (1997, p.1) defines it as "a body of naturally-occurring language (authentic) data", while Kennedy (1998, p. 1) puts it as a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description". Similarly, O'Keeffe, McCarthy and Carter (2007, p.1) indicates that corpora (plural for *corpus*) are "principled" collection of texts, and not any collection can be considered as corpus; a corpus has to be representative.

It goes without saying that today almost all corpora appear in electronic form. With the advent of computer technology, the ability of searching for, retrieving, sorting and calculating data have provided a valuable aid to linguists (Leech, 1991). One can search for a word to see its usage, retrieve examples in context, calculate the frequency of its occurrences, and sort the data in some way, such as alphabetically. These abilities of "machine readable" corpora (McEnergy & Wilson, 2001) render both the collection and analyzes of texts faster and more reliable.

Today corpora are available for many languages, representing the language in general terms as well as its different varieties. They can contain only spoken or written texts, or a combination of the both. Regarding their purpose, Hunston (2002) classifies corpora as ‘specialized corpus’, ‘general corpus’, ‘comparable corpora’, ‘parallel corpora’, ‘learner corpus’, ‘pedagogic corpus’, ‘historical or diachronic corpus’ and ‘monitor corpus’. Two of them were employed in the present study; a general and pedagogic corpus. A pedagogic corpus, as Flowerdew (2012) states, is “a corpus which has been compiled from the language teaching textbook(s) to which the students has been exposed” (p. 323). In order to see to what extent it includes authentic language, most of the time it is compared with a general corpus, which includes texts of many types of written and/or spoken language, produced in one country or many. A general corpus is generally used with the aim of producing reference materials for language learning; therefore, it is also called a ‘reference corpus’. Table 2.1 is a summary of frequently referred and freely available three general corpora, which are the British National Corpus (BNC), containing about 100 million words; the Corpus of Contemporary American English (COCA), containing over 450 million words; and the Open American National Corpus (OANC), containing 15 million words.

Table 2.1
Summary of Freely Available General Corpora

Feature	COCA	BNC	OANC
Availability	Free / online	Free / online	Free / downloadable
Size (millions of words)	450	100	15
Time span	1990-2012	1970s-1993	2000-2005
Number of words of text being added each year	20 million	0	0
Can be used as a monitor corpus to see ongoing changes in English	Yes	No	No
Wide range of genres: spoken, fiction, popular magazine, newspaper, academic	Yes	Yes	No
Size of spoken (millions of words)	90	10	4
Dialect	American	British	American

Note. Adapted from the official web page of the Corpus of Contemporary American English (COCA), http://corpus.byu.edu/coca/

2.2. The Indirect Use of Corpus and What it can Offer in Language Teaching Materials

Basically, there are two ways of utilizing corpus in classifying pedagogical applications; *direct* and *indirect* (see Figure 2.1) (Römer, 2011; Stubbs, 2004).

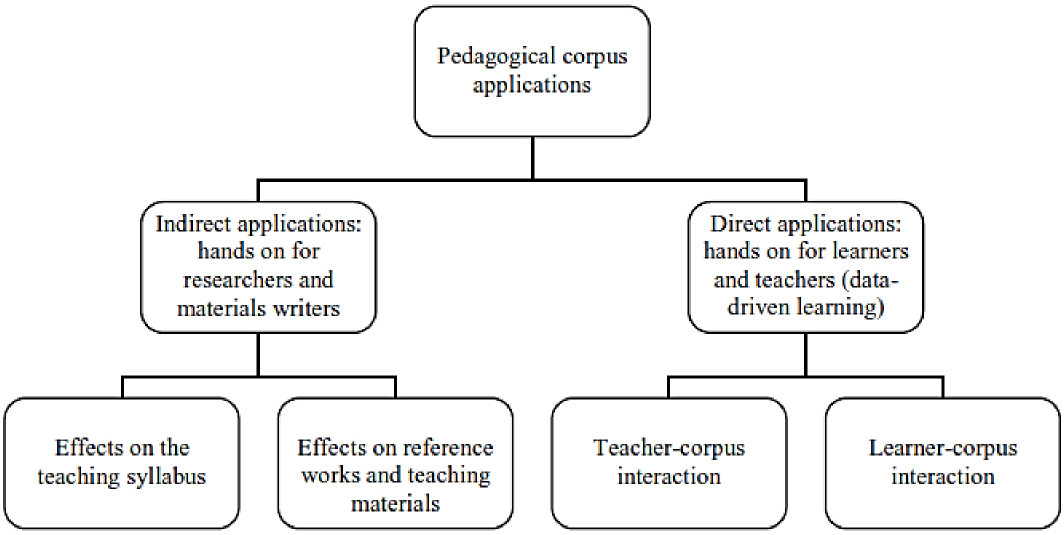

Pedagogical corpus applications
Indirect applications: hands on for researchers and materials writers
Direct applications: hands on for learners and teachers (data-driven learning)
Effects on the teaching syllabus
Effects on reference works and teaching materials
Teacher-corpus interaction
Learner-corpus interaction

Figure 2.1. The use of corpora in second language learning and teaching (Römer, 2011).

According to Leech (1997), the direct way includes 'teaching about', 'teaching to exploit', and 'exploiting to teach'. While the first one means teaching corpus linguistics as an academic subject, the other two is about how to use corpus, which can be expected to be seen in advanced levels (McEnergy and Xiao, 2010). McEnergy and Xiao also argue that because of several reasons such as the level and experience of learners, the restrictions in the access to necessary electronic resources and the lack of teacher skill for corpus analysis, the use of corpora in language teaching and learning has been more indirect than direct. The indirect use of corpus in language teaching/learning includes such areas as reference publishing, materials evaluation and development, and testing.

A corpus provides its users with a real picture of how people speak or write in specific contexts. Through databases containing millions of words, one can retrieve information about the use of vocabulary, grammar and discourse, as well as highlighting the differences between the spoken and written languages. Moreover, textbook writers can utilize the statistics a corpus can provide as it is easier to make analyzes on real language output by using corpus software. Lexical and grammatical profiles of a language for specific purposes, for example, can be identified to gain a deeper insight about how authentic language works (O'Keefe, McCarthy, & Carter, 2007; Schmitt, 2010).

Employing corpus methodology in textbook writing process helps authors rely on empirical data rather than their intuitions while designing the textbook. Moreover, McCarthy (2004) argues that students' knowing that the materials they are studying are based on authentic use of the language boosts their motivation in learning; consequently, corpus-informed materials take them to the target more efficiently. He also points out the advantages of corpus-informed language materials as the following (McCarthy, 2004, p.15):

- The examples used in them, although they may sometimes be edited or adapted, are a reflection of real usage; they are not invented.
- The syllabus (the items to be taught as well as the sequence in which they will be presented) is informed by frequency information: For instance, we can prioritize grammar and vocabulary that is most frequent and most useful.
- The contexts in which words and grammar structures are used are authentic ones, based on the contexts that occur in corpora.
- The presentation and activities can focus on the important differences between spoken and written language.
- The materials can include language that was ignored or not noticed in the past but that is at the heart of real communication.
- Specialized corpora can be analyzed to meet the needs of particular groups of learners. For example, we can use an academic corpus collected in university and college

contexts to help learners who are going to study abroad, or a business corpus to construct materials for businesspeople who need to work in a second language.

- The writers of corpus-informed materials can anticipate common errors by looking at corpora of learners' work from a wide variety of language backgrounds.
- Students don't have to live in the target language environment to experience authentic language – it's right there, in their course books and dictionaries.

2.3. Textbook Evaluation

Obviously, there is a wide range of English textbooks available in the market. In order to pick up one among them, teachers and educational administrators need to undertake an evaluation process. In this process, making comparisons is inevitable to weigh the advantages and limitations of the available textbooks. Although evaluators might have some precious insights about teaching, ideally they are expected to be as objective as possible and base their selection on empirical data. However, it is very challenging to accomplish a thorough textbook analysis manually. At this point corpus linguistics provides an easier and more objective alternative for quantitative textbook evaluation.

Textbook evaluation via corpus enables researchers and evaluators to compile textbook corpus, and compare and contrast textbooks with each other as well as with a native speaker corpus, focusing on various aspects of the language. Thanks to the development in the computer technology, several searches can be made using a concordancer (a software program to search through a corpus) in a relatively short time. The data retrieved can be benefited in many aspects in ELT, including textbook evaluation.

As part of a project, Ljung (1991) compares the 1,000 most frequent words in the GYM corpus (a corpus of textbooks studied in gymnasium) and COBUILD as a reference corpus in order to evaluate the textbooks at vocabulary level. The comparison is based on the unique words and differences in frequency between the shared words. The study reveals that the words found only in the GYM corpus are concrete terms and of narrative kind, while the words unique to the COBUILD corpus are predominantly abstract. In 1999, Ljung also concludes at the end of the project that GYM textbooks are insufficiently progressive as the difficult words were very often randomly distributed across the grades, instead of growing successively from more common to less frequent.

Super (2004) suggests that corpora can be of help in ELT in three aspects: textbook evaluation, textbook development and ESP materials development. She reports on the findings of a study, which included a comparison of frequencies of expressions and collocations in two corpora: an academic speech textbook – Discussion and Interaction in the Academic Community (Madden & Rohlck, 2000), and Michigan Corpus of Academic Spoken English (MICASE). The study revealed that the textbook did not actually represent authentic academic speech, and it went through an update process so that it could include more realistic content. Several other researchers have used corpus to look critically at the existing teaching materials. Before reviewing more studies, it would be more appropriate to discuss what *authenticity* means in ELT textbooks.

2.4. Authenticity and 'Textbook English'

Gray (2002) states that the sales of course books reach hundreds of thousands a year, which implies that course books have a great impact on both teachers and students throughout the world. However, some researchers looking critically at existing TEFL materials argue that they are not authentic. In this context, according to McDonough and Shaw (2003) authenticity is "a term that loosely implies as close an approximation as possible to the world outside the classroom, in the selection both of language material and of the activities and methods used for practice in the classroom" (p. 41).

McEnergy and Xiao (2010) argue that learners find it difficult to speak English. One of the solutions to this problem can be to expose students to more authentic examples of spoken language, with which corpora can provide us. Sometimes, inadequate or faulty examples of language cause learners to be unable to learn authentic English. In 1998, Carter gathered a corpus of dialogues from textbooks and compared it to the CANCODE (Cambridge and Nottingham Corpus of Discourse in English) spoken corpus. The study showed that the textbook dialogues lacked many spoken language features. Course books' containing imaginative and intuitive context both lexically and grammatically may be considered as a disservice to students.

The mismatch between the English presented in course books and the English used in real life by native speakers has been pointed out in several studies (Klages & Römer, 2002; Römer, 2004a, 2004b, 2005a, 2005b). Referring to this mismatch, Römer names the English presented in course books as 'textbook English'. In one of her studies (2004a), for example, she compiled a German EFL spoken textbook corpus, representing the speech production in the selected textbooks, and compared it with the spoken part of BNC (British National Corpus) focusing on the use of if-clauses. She argues that the if-clause representations in those textbooks are not accurate and do not include all the variations in real life. In another study (2005b), analyzing the use of progressive forms in German EFL textbooks in comparison with two native spoken corpora, she reports a discrepancy between the two. Similarly, Mindt (1996) indicates that the use of grammatical structure differs considerably from native speakers' use. That is why he argues that students who have been taught 'school English' have difficulties in dealing with real life English.

While a great number of researchers argue for the value of using real language in textbooks (Davidson, Indefrey, & Gullberg, 2008; Harwood, 2005; O'Keefe et al., 2007; Römer, 2004a), there are some who suggest that authenticity is impossible to accomplish because of the non-authentic nature of the EFL classrooms (Cook, 2001; Widdowson, 2000). They argue that the way that native speakers use the language cannot be represented in classroom environment, which is not natural, and using invented examples in textbooks is helpful in teaching specific structures.

Having reviewed the literature about using corpora in ELT and providing learners with authentic materials, in the following paragraphs, I will point out the importance of frequency data, which can be retrieved via corpus software, and elaborate on vocabulary frequency as well as four-word bundles frequency.

2.5. Frequency

Biber and Reppen (2002) suggest that textbook writers and language learners have always been interested in the lists of phrasal verbs and other expressions that are commonly used. As an authentic source of language, corpora can be used to provide frequency data of the language, which may be helpful in language materials design. The data can show EFL textbook writers, teachers and learners the differences between written and spoken corpora, native and non-native speaker corpora regarding the occurrences of specific words and multi-word combinations.

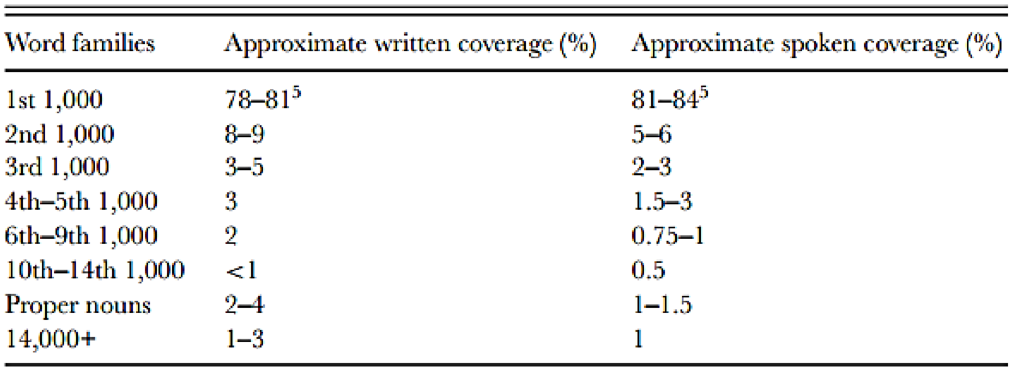
Mindt (1996) made a comparison of the frequencies of modal verbs, future time expressions and conditional clauses in the BNC (British National Corpus) and their grading in textbooks in German context, and she argues that frequency of usage can be a guide while designing textbooks and grading them. Rather than basing the content on the intuitions of the author, it is wise to trust empirical data that is provided from corpora of native speakers. The frequencies of individual words, modal verbs and tenses can be taken as a base in the design. Goethals (2003) also recognizes the importance of the frequency information in grading language items in textbooks as he considers the information as "a measure of *probability* of usefulness" (p. 424). Though not the only factor language pedagogy should be counted on, frequency information is highly valued in language learning/teaching. Leech (1997) argues that "Whatever the imperfections of the simple equation 'most frequent' = 'most important to learn', it is difficult to deny that frequency information becoming available from corpora has an important empirical input to language learning materials" (p. 16).

2.6. Vocabulary Frequency

As McCarthy and Carter (2002) suggest, single words are considered as the central units to be acquired while learning English. The frequency distribution of the vocabulary is significant in the teaching of lexis since it brings the learners closer to native speaker norms (Shirato & Staplaton, 2007). Consequently, many researchers agree that high-frequency vocabulary must be taught explicitly in foreign language classes (Nation, 2001; Schmitt, 2011). It is essential to find out how frequently language items occur frequently in specific contexts as well as to enable a comparison between different situations. This way, it is possible to explore the characteristics and densities of written and spoken texts in addition to providing learners with a list of words for instructional and informative purposes (Gardner, 2007).

There are few students who have a chance to experience the language in its natural environment. Many language learners are not exposed to any language input other than textbooks. Thus, as the major source of language, it is important to investigate the lexical content and authenticity of EFL textbooks.

Although it is not realistic to expect learners to learn all the vocabulary in a textbook, researchers argue that frequently occurring words are likely to be known (Kachroo, 1962; Saragi, Nation, & Meister, 1978; Zahar, Cobb & Spada, 2001). They are regarded to be a significant aspect of the language in terms of cost/benefit; consequently, they are advised to be taught explicitly (Nation & Waring, 1997; Shmitt, 2011). Based on Table 1 from Nation's study (2006, p. 79), Schmitt and Schmitt (2012, p. 5) suggest that high-frequency vocabulary would include the most frequent 2–3,000 word families in English. Knowing them would help the learners understand a great deal of written or spoken language.



Word families	Approximate written coverage (%)	Approximate spoken coverage (%)
1st 1,000	78–81[5]	81–84[5]
2nd 1,000	8–9	5–6
3rd 1,000	3–5	2–3
4th–5th 1,000	3	1.5–3
6th–9th 1,000	2	0.75–1
10th–14th 1,000	<1	0.5
Proper nouns	2–4	1–1.5
14,000+	1–3	1

Figure 2.2. Vocabulary size and text coverage (written and spoken) across nine spoken and written corpora (Nation, 2006, p. 79).

Researchers have been interested in examining textbooks' vocabulary content (Biber, Conrad, & Cortes, 2004; Meunier & Gouverneur, 2007; Matsuoka & Hirsh, 2010). Matsuoka and Hirsh (2010), for example, examined the vocabulary use in an upper-intermediate course book. They argue that the course book offers few opportunities to expand vocabulary knowledge beyond the most frequent 2,000 words, and 33.3% of the 2k words occurring in the textbook are repeated only once. Martini (2012) reports that Horst, White, and Cobb's paper presented in a conference found that about 26% of the 1k, 2k, and 3k levels were missing in the textbook corpus of a typical series of Quebec primary ESL textbooks.

2.7. Multi-Word Combinations or Lexical Bundles

Many researchers (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Hakuta, 1974; Nattinger & De Carrico, 1992; Wray, 2000, 2002) have studied the notion of multi-word combination under many rubrics including lexical phrases, fixed expressions, formulas, phraseological units, routines, lexical bundles, prefabricated patterns, formulaic sequences, chunks, and recurrent word combinations. Choosing among these, the present study focuses on the notion designated by the term 'lexical bundles' coined by Biber *et al.* (1999). The term *lexical bundles* have distinct pragmatic functions in spoken and written discourse (Chen, 2010). Biber *et al.* (1999) defined lexical bundles as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (p. 990). In Biber and Conrad's study (1999) the term is defined as "the most frequent recurring lexical sequences; […] which can be regarded as extended collocations: sequences of three or more words that show a statistical tendency to co-occur (e.g., *in the case of the*)" (p. 183).

According to Biber, Conrad and Cortes (2004) much of our everyday language use is composed of prefabricated expressions, and there is a general consensus on the importance of lexical bundles. Recently, several research studies on lexical bundles have been conducted (Biber & Barbieri, 2007; Cortes, 2004; Cortes, 2008; Kim, 2009). Jalali and

Ghayoomi (2010) argue that the reason why the term 'lexical bundles' attracts researchers' attention is "their functional contribution to the coherence and organization of different texts, either spoken or written, rather than their pervasive presence" (p. 324). In both spoken and written language, recurrent structures are found to be very common (Schmitt, 2012). According to Biber *et al.* (1999), around 30% of the words in their conversation and 21% of the words in their academic prose corpus are made up of lexical bundles. Regarding the functional classification of these word combinations, Biber et al, (2004) suggest that they serve for a wide range of discursive functions such as organization of discourse, expression of stance, and reference to textual or external entities.

Several research studies have shown that the frequency of multi-word combinations surpass that of single words (McCarthy & Carter, 2002; Altenberg & Granger, 2001), which indicates that they are as crucial as single words in communication. Without doubt, they function to decrease hesitation and pauses in speech. Partington (1998) points out their importance in communication for both the hearer and the speaker stating that "language consisting of a relatively high number of fixed phrases is generally more predictable than that which is not" (p. 20). However, De Cock (1998) contends that they were neglected in language teaching because their value was not well-recognized.

On the other hand, researchers who recognize their value need to limit their study. It is such a broad area to conduct a study on all lexical bundles in a corpus. Consequently, researchers conducting a corpus-based research on lexical bundles tend to narrow it down to make it more feasible, by focusing on the ones with a specific number of words. Yet, as Jablonkai (2009) argues that four-word bundles are more frequent compared to others, they give more variety for the structural and functional analysis than five-word bundles, and three-word lexical bundles are often part of four-word bundles. Moreover, Biber et al. (1999) suggest that four-word bundles and above "are more phrasal in nature and correspondingly less common" (p. 992). This may be the reason why several researchers have focused on four-word lexical bundles (Chen & Baker, 2010; Hyland, 2012; Jablonkai, 2009).

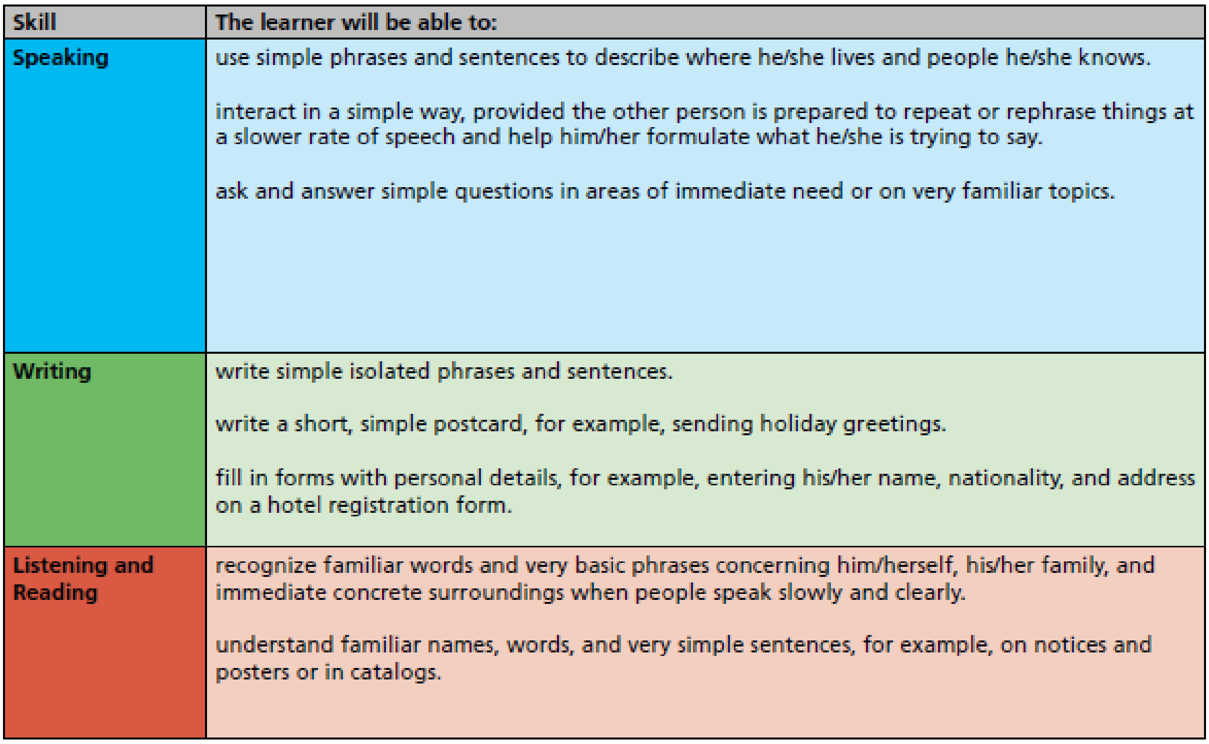
III. METHODOLOGY

3.1. Materials

The present study needed three corpora, a corpus of EFL textbooks published in Turkey by various Turkish publishers, a corpus of a corpus-informed EFL textbook by a respected publisher and a reference corpus. To compile the textbook corpus, 11 English course books together with their workbooks and teacher's books, which are approved by the Ministry of National Education MoNE in Turkey to be taught at state primary and secondary schools in Turkey, were employed. Those textbooks were published either by the MoNE itself or by the private publishers which cooperate with the MoNE. In those textbooks, the topics and the functions aimed to be realized in each unit were predefined by the MoNE, and regardless of the publisher, and they were designed accordingly. Grade 4 and 5 textbooks included 14 chapters while the Grade 6, 7 and 8 textbooks included 16 chapters about the same topics aiming to teach the same grammar functions. For a grade, there are more than one alternative to be employed as a textbook.

The education system in Turkey changes very often, and is different currently. At the time of the present study, English was taught as a foreign language at state schools starting from Grade 4 to Grade 8, the end of secondary education. The basic and compulsory education in Turkey consisted of 8 years of education, which means students in Turkey have to be a secondary school graduate at least. During this education, students at state schools started to learn English at Grade 4 (approximately at the age of 10) until the graduation at the end of Grade 8 (approximately at the age of 14). When students first start to learn English at Grade 4, it could be assumed that they are zero beginners, unless they have already attended private English courses previously. However, it is not clear which level students are expected to reach by means of the English language education provided at state schools in Turkey. There is no information regarding the target English level of students at the end of the compulsory education neither on the website of the MoNE.

In addition to the textbooks approved by the Turkish MoNE, a corpus-informed textbook, *Touchstone 1*, was used to answer the second research question. *Touchstone* series is based on the North American English portion of the Cambridge International Corpus (currently known as Cambridge Corpus of English), and *Touchstone 1* is the lowest level of the series, covering level A1 of the Common European Framework of Reference for Languages (CEFR). Figure 3.1 describes the general degree of skill achieved by learners at this level (Cambridge English CEFR Correlations). Based on this description, it was supposed that *Touchstone 1* is the counterpart of the textbooks of Grades 4 and 5.



Skill	The learner will be able to:
Speaking	use simple phrases and sentences to describe where he/she lives and people he/she knows. interact in a simple way, provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help him/her formulate what he/she is trying to say. ask and answer simple questions in areas of immediate need or on very familiar topics.
Writing	write simple isolated phrases and sentences. write a short, simple postcard, for example, sending holiday greetings. fill in forms with personal details, for example, entering his/her name, nationality, and address on a hotel registration form.
Listening and Reading	recognize familiar words and very basic phrases concerning him/herself, his/her family, and immediate concrete surroundings when people speak slowly and clearly. understand familiar names, words, and very simple sentences, for example, on notices and posters or in catalogs.

Figure 3.1. The description of the general degree of skill achieved by learners at A1 level according to CEFR.

According to McCarthy (2004), one of the authors of the series, the most common words in the order of their frequency were employed in the textbook after a study on the Spoken and Written Corpus. Moreover, he argues that the content is authentic and based on the current usage of the language to communicate in everyday situations, especially in conversation.

As for the reference corpus, the spoken component of the Open American National Corpus (OANC) was preferred (referred as *the OANC-spoken* in the present study). The OANC-spoken includes 3,217,772 words from face-to-face and telephone conversations of hundreds of American English native speakers (for more information about the OANC-spoken, visit http://www.anc.org/data/oanc/contents/). There are three reasons why the OANC-spoken was used in this study. First, as several researchers suggest (Greenbaum, 1990; Matsuda & Friedrich, 2012), American English is one of the well-accepted varieties of English, and the OANC is a corpus representing this variety. Second, it includes spoken data representing a widely accepted variety of English; American English. Lastly, it is made available to researchers at no cost to download.

In order to retrieve the frequency lists from both the pedagogical corpora and the reference corpus, a software program called *concordancer* was needed. AntConc (Anthony, 2011) developed by Laurence Anthony was chosen because it is user-friendly and available for download at no cost.

3.2. Procedures

3.2.1. Corpus design

In the present study, the term ‘textbook’ is used to refer to a compilation of a course book (i.e. student book) (SB), workbook (WB) and teachers’ book (TB). Each course book is accompanied by a workbook and a teachers’ book. As the main focus of the study is on how the textbooks present the spoken English, only the listening and conversational parts; i.e. dialogues, speech bubbles, and songs in both course books and workbooks, were included as well as the audio scripts provided in the teachers’ books. Moreover, the answer keys, if provided in the teachers’ books, were also taken into account for the missing parts in the exercises having a conversational aspect.

When considered as a whole, the conversational parts of 33 books were compiled to make up the corpus of ELT textbooks for Grades 4, 5, 6, 7 and 8, which are approved by the MoNE (TC-Tr). Moreover, the course book, workbook and teachers' book of *Touchstone 1* were studied in the same way to compose the *Touchstone 1* corpus (TC-Ts). For this purpose, the pedagogical corpora; i.e. the TC-Tr and the TC-Ts, were gathered in PDF form, and converted into plain text so that it could be possible to analyze the data on AntConc. Then, the texts were cleaned out to exclude the parts other than the conversations and the audio scripts.

After compiling the pedagogical corpora, the texts were cleaned out of all punctuation marks as they are not concrete in speaking. Since they were replaced with blank, contractions such as *don't*, *isn't* and *can't* were considered as two words, counting *t* as a word. Then, all the course books, workbooks and teachers' books were considered in a group of their own, and the one which had the fewest word count was taken as a base in each group, which is 1197 words for SBs, 365 words for WBs and 1308 words for TBs. The purpose was to obtain equal samples from each book to enable a comparison between homogeneous groups. Then, each equalized course book, workbook and teachers' book were gathered under the textbook plain text document they belong to. As a result, the word count for each textbook researched were equal, which is 2870 in total (see Table 3.1).

Table 3.1

The equalized word counts for each SB, WB and TB.

Name	Publisher	Type	Grade	Word Count
1. Joyful English 4	MoNE	SB	4	1197
2. Joyful English 4	MoNE	WB	4	365
3. Joyful English 4	MoNE	TB	4	1308
4. English 4	SEK	SB	4	1197
5. English 4	SEK	WB	4	365
6. English 4	SEK	TB	4	1308
7. Time for English 5	MoNE	SB	5	1197
8. Time for English 5	MoNE	WB	5	365

9. Time for English 5	MoNE	TB	5	1308
10. My English 5	Pasifik	SB	5	1197
11. My English 5	Pasifik	WB	5	365
12. My English 5	Pasifik	TB	5	1308
13. Spot On 6	MoNE	SB	6	1197
14. Spot On 6	MoNE	WB	6	365
15. Spot On 6	MoNE	TB	6	1308
16. Unique 6	Atlantik	SB	6	1197
17. Unique 6	Atlantik	WB	6	365
18. Unique 6	Atlantik	TB	6	1308
19. Spot On 7	MoNE	SB	7	1197
20. Spot On 7	MoNE	WB	7	365
21. Spot On 7	MoNE	TB	7	1308
22. Texture English 7	Doku	SB	7	1197
23. Texture English 7	Doku	WB	7	365
24. Texture English 7	Doku	TB	7	1308
25. Spring 7	Ozgun	SB	7	1197
26. Spring 7	Ozgun	WB	7	365
27. Spring 7	Ozgun	TB	7	1308
28. Spot On 8	MoNE	SB	8	1197
29. Spot On 8	MoNE	WB	8	365
30. Spot On 8	MoNE	TB	8	1308
31. Four Seasons English 8	Dikey	SB	8	1197
32. Four Seasons English 8	Dikey	WB	8	365
33. Four Seasons English 8	Dikey	TB	8	1308
34. Touchstone 1	Cambridge U. P.	SB	A1 (4-5)	1197
35. Touchstone 1	Cambridge U. P.	TB	A1 (4-5)	1308
36. Touchstone 1	Cambridge U. P.	WB	A1 (4-5)	365

3.2.2. Frequency analysis

Linguists have different suggestions as to which words to be considered among *high-frequency* vocabulary. Some studies refer to the most frequent 2000 word families as high-frequency vocabulary (Schmitt, 2000; Nation, 2001; Thornbury, 2002). However, in a more recent study, Schmitt and Schmitt (2006) suggest that the most frequent 3000 word families should be recognized as the high-frequency vocabulary in English. Considering that the levels of the textbooks employed in the present study were thought to be low (A1 and A2, although not cited by their authors), it was taken for granted that the textbooks did not include many word families. At this point the distinction between lemmatized and non-lemmatized forms of the words, and word families should be clarified. Biber (2006, p. 242) refers to *lemmas* as the base forms of each word, disregarding inflectional morphemes. Consequently, *work*, *works*, *worked* and *working* are taken as realizations of a single lemma; *work*. On the other hand, word families "include 'closely related derived forms' in addition to all inflected variants for a word" as reported from Nation's (2001, p. 8) study by Biber (2006); as a result, *work*, *works*, *working*, *worked*, *workable*, *worker* are all considered to be one type and members of the same word family. Considering the classification problem of some word families due to their multiple word-class membership, in the present study each word was used as the basic unit of the analysis, regardless of its being derived or inflected from a certain root or its belonging to a word family.

3.2.2.1. Reference list of words

In corpus studies, word frequency is often described in 1000 bands, and referred to as 1k for the first 1000 words (Nation, 2001; Schmitt & Schmitt, 2012). The present study also adopts this style, and refers to the second and third 1000 bands as 2k and 3k. In order to answer the first parts of the research questions, the most frequent 3000 reference words in OANC-spoken were identified in three groups consisting of a thousand words (represented as 1k, 2k and 3k words) in each (see Appendix A1), and saved in plain text documents.

AntConc was used to create a frequency list of single word items out of the OANC-spoken, which consists of 3,251,951 running words (tokens) and 28,217 word forms (types). The function *Word List* was selected, and the tool was set to analyze all data in lower case. By this way, the 1k, 2k and 3k bands were retrieved to be checked in both the TC-Tr and the TC-Ts, and saved them separately in 1000-word sets in *.txt* extensions. To get the results, each textbook set, which consists of a student book, a workbook and a teachers' book, were uploaded to AntConc software one by one, and for each textbook Advanced Search option was used in order to check the occurrences of each band (see Appendix A2, for a textbook analysis). However, AntConc could only give the tokens, so the results were sorted alphabetically and the recurrent words were manually counted as one item to get the type count.

3.2.2.2. Reference list of four-word bundles

Similar steps were carried out to answer the second parts of the research questions, which deal with the authenticity of the textbooks considering four-word bundles. The tool was again set to analyze all data in lower case. Then, the Cluster analysis was conducted with "n-gram" command set at 4-grams. By this way, the most frequent 3000 four-word-bundles (represented as 1k, 2k and 3k four-word bundles) in the OANC-spoken, which includes 238,374 four-gram tokens and 4535 4-gram types, were analyzed.

According to Biber (2006) and Flowerdew (2012), the cut-off points chosen to count recurrent combinations as bundles is often arbitrary. While Biber et al. (2004) considers the four-word lexical combinations occurring 40 times or more per million words as bundles, Cortes (2004) concludes that number should be 20 or more per million words. The present study adopted a less conservative approach, and set the cut-off point at 20 (see Appendix B1). However, the minimum 4-gram frequency in the most frequent 3000 was found to be 26. Then, the results for each band were saved in a plain text document. Retrieving the results, their occurrence was checked by uploading each a thousand-set in Advanced Search. Textbooks were loaded one by one for each batch search so that they can be checked at once, by using Concordance search section (see Appendix B2, for a textbook analysis). Again, as it was not possible to have the type counts automatically on AntConc, this was done manually for each band and each textbook after sorting the concordance lines alphabetically. To achieve this, KWIC (Key Word in Context) sort option was used and set at 0 for Level 1, 1R for Level 2 and 2R for Level 3 (see Appendix C).

IV. RESULTS

The first research question examined the extent to which the corpus of ELT textbooks approved by Turkish Ministry of National Education (MoNE) for grade 4, 5, 6, 7 and 8 (TC-Tr) mirror the authentic spoken English, regarding the most frequently used 1st, 2nd and 3rd 1000 single word and four-word bundle bands. Whereas the second research question sought to examine how well the corpus of *Touchstone 1* (the TC-Ts), a corpus-informed textbook, reflect the authentic spoken English in comparison with the other textbooks from the TC-Tr. The reference list obtained from the OANC-spoken were organized into both single word (see Appendix D) and four-word bundle (see Appendix E) bands as 1k, 2k and 3k for each 1000 set. This kind of an organization made the comparison of the TC-Tr and the TC-Ts easier; consequently, the data are analyzed accordingly in two groups; vocabulary and four-word bundles in the textbook corpora. As a result of the search carried out using the lists from the OANC-spoken, it was discovered how many of these words/bundles were contained in each of the textbooks.

4.1. Comparing the Use of Single Words from the Reference Lists

Table 4.1 gives the number of types and tokens of the reference words which the TC-Tr and the TC-Ts include from the bands and in total. At the 1k level, *Texture English 7* has the highest number of types from the reference list (with 451), while the highest frequency is in *Spring 7* with 2,485 tokens. However, the number of the tokens in the TC-Ts (2594) is even more than the one in *Spring 7* has, which means that 90.4% of words is made up of the words in the 1st 1000 word list (Table 4.2). At the 2k level, *Spot On 6* has 113 of 1000 words, which is the highest number in the TC-Tr. As for the number of tokens, the highest value is in *English 4* with 300, making 10.5% of words it includes. On the other hand, only 4.5% of the TC-Ts is formed by the words in the 2nd 1000 word list (Table 4.2). As far as the 3rd 1000 word band is concerned, again *Spot On 6* includes the highest number of types, while *Time For English 5* is the one with the highest number of tokens with 132. Although *Joyful English 4* has only 22 words from the band, it is the second textbook in the TC-Tr recycling those words mostly in 2870 words, with 4.2%, after *Time For English 5* (Table 4.2). In TC-Ts, there are 36 word types, and they are used 85 times, which means 3% of the whole corpus.

Table 4.1

Numbers of types and tokens for the TC-Tr and the TC-Ts at 1k, 2k and 3k levels in single word inquiry.

Textbook	Grade	1k		2k		3k		Total	
		Type	Token	Type	Token	Type	Token	Type	Token
1. Joyful English 4	4	158	2371	49	176	22	121	229	2668
2. English 4	4	211	2304	70	300	45	110	326	2714
3. Time for English 5	5	167	2287	79	209	56	132	302	2628
4. My English 5	5	246	2283	90	199	48	111	384	2593
5. Spot On 6	6	339	2242	113	204	64	116	516	2562
6. Unique 6	6	302	2421	96	245	48	89	446	2666
7. Spot On 7	7	314	2301	74	194	58	118	446	2613
8. Texture English 7	7	451	2380	89	179	51	89	591	2648
9. Spring 7	7	284	2485	93	163	52	83	429	2731
10. Spot On 8	8	392	2306	103	194	58	99	553	2599
11. Four Seasons English 8	8	346	2382	82	158	55	99	483	2639
12. Touchstone 1	4-5	279	2594	45	130	36	85	360	2809

As can be seen in the last column in Table 4.1, in the TC-Tr *Texture English 7* is the textbook having the highest number of word types from the most frequently used 3000 words in OANC-spoken, with 591 words. Concerning the number of word tokens, *Spring 7* is at the top with 2731 tokens in total from the 3000 words, which makes 95.2% of the total word number as indicated in Table 4. On the other hand, in the TC-Ts there are 360 word types and 2809 tokens in total. This means that the words from 360 words from the most frequently used 3000 words in the OANC-spoken are recycled 2809 times, constituting 97.9% of the whole corpus, which is more than any textbook in the TC-Tr does.

Table 4.2

Percentages showing how much of dialogues in each textbook is composed of the words from the reference bands.

Textbook	Grade	1k	2k	3k	Total
1. Joyful English 4	4	82.6	6.1	4.2	92.9
2. English 4	4	80.3	10.5	3.8	94.6
3. Time for English 5	5	80.3	7.3	4.6	92.2
4. My English 5	5	79.7	6.9	3.9	90.5
5. Spot On 6	6	78.1	7.1	4	89.2
6. Unique 6	6	84.3	8.5	3.1	95.9
7. Spot On 7	7	80.1	6.8	4.1	91
8. Texture English 7	7	82.9	6.2	3.1	92.2
9. Spring 7	7	86.6	5.7	2.9	95.2
10. Spot On 8	8	80.3	6.8	3.4	90.5
11. Four Seasons English 8	8	83	5.5	3.4	91.9
12. Touchstone 1	4-5	90.4	4.5	3	97.9

Table 4.3 presents the numbers in Table 4.1 in percentages so that the comparison of the textbooks could be clearer considering their vocabulary coverage from the bands. The percentages of the 2nd and 3rd vocabulary bands covered in the TC-Tr and the TC-Ts decline dramatically compared to the ones for the 1st band. It can be seen that *Texture English 7,* which has the highest number word types in the 1k band, includes 45.1% of the most frequent 1st 1000 words in the OANC-spoken. However, it drops to 8.9% and 5.1% when the words from 2k and 3k bands are checked, respectively. One can see a decline in the TC-Ts, as well. The percentage of the 1k words from the OANC-spoken occurring in the TC-Ts, 27.9%, drops to 4.5% and 3.6% at 2k and 3k bands, respectively. Figure 4.1 also visualises the amount of the words covered from the whole reference list in percentage.

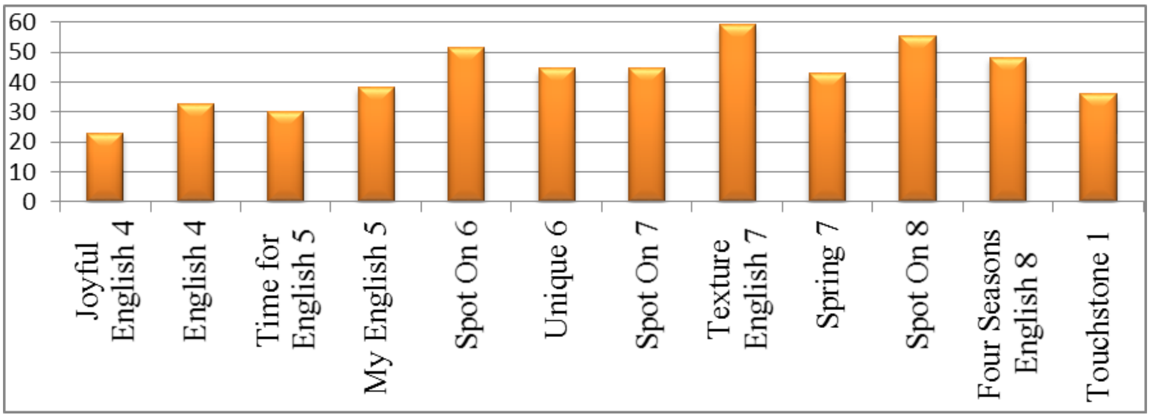


Figure 4.1. A comparison of the textbooks regarding the percentage of the words covered from the whole reference list.

In Table 4.3, it can also be observed that there is no gradual increase in the percentages of higher band words when moved to a higher grade. On the contrary, there is even a decrease compared to a previous grade's percentages in some cases. For example, *English 4* includes 21.1% of the most frequent 1st 1000 words in the OANC-spoken. In Grade 5, although *Time For English 5* has 16.7% of those words, *My English 5* has 24.6% of them. In Grade 6, both alternatives, *Spot On 6* (33.9%) and *Unique 6* (30.2%) have a higher percentage for 1k band compared to the previous grade's percentages. There are three alternative textbooks in Grade 7, some of which have a higher percentage of words from 1k band, depending on the one chosen. This is also the case for Grade 8. In other words, one can observe both an increase and a decrease in the percentages compared to the previous grade's percentages, depending on the textbook studied. Consequently, it is not possible to give a definite number for word types covered in the textbooks of the same grade.

Moreover, even among *Spot On* series, which has textbooks for Grade 6, 7 and 8, there is no consistent increase in the percentages of the words included from the most frequent 3000 words in the OANC-spoken. While 33.9% of the words in 1k band occur in *Spot On 6*, 31.5% of them can be seen in *Spot On 7*. Then, in *Spot On 8*, the percentage rises up to 39.4%. On the other hand, it is not possible to compare *Touchstone* series textbooks in itself as in the present study only *Touchstone 1* was analyzed.

Table 4.3

Percentages of the vocabulary bands covered in the TC-Tr and the TC-Ts.

Textbook	Grade	1k	2k	3k	Total
1. Joyful English 4	4	15.8	4.9	2.2	22.9
2. English 4	4	21.1	7	4.5	32.6
3. Time for English 5	5	16.7	7.9	5.6	30.2
4. My English 5	5	24.6	9	4.8	38.4
5. Spot On 6	6	33.9	11.3	6.4	51.6
6. Unique 6	6	30.2	9.6	4.8	44.6
7. Spot On 7	7	31.4	7.4	5.8	44.6
8. Texture English 7	7	45.1	8.9	5.1	59.1
9. Spring 7	7	28.4	9.3	5.2	42.9
10. Spot On 8	8	39.2	10.3	5.8	55.3
11. Four Seasons English 8	8	34.6	8.2	3.6	48.3
12. Touchstone 1	4-5	27.9	4.5	3.6	36

4.1.2. Reference words that were never used in the textbooks

In order to see some examples of the words from the lists that occur and do not occur in the textbook corpora, a random search was carried out for all bands. Words such as “stuff”, “definitely” and “somewhere” were found to be non-existent in the textbook corpora in the search for the first band. On the other hand, “sort” exists only in *Spring 7* while “well” exists in all the textbooks except for *English 4*. In the 2nd band search “somehow” and “weird” were found to be absent in the textbooks. However, “environment” occurs in *Spot On 7* and *Spring 7*; “busy” occurs in *Touchstone 1*, *Spot On 6, Unique 6* and *Four Seasons English 8*; and “tonight” occurs only in *Spring 7*. Finally, in the last band search, “appropriate”, “seriously” and “anyhow” were not found in the textbooks while “opposite” exists in *Unique6*, *Spot On 7* and *Spring 7*, and “directly” exists only in *Four Seasons English 8*.

4.2. Comparing the Use of Four-word Bundles from the Reference Lists

Comparing the use of four-word bundles from the reference lists in Table 4.4, one can realize that few of them were covered in the textbooks. Out of 1000 in each band, the highest number of type is 24, in *Spot On 7* for the 1st 1000. Token-wise, as can be expected, the highest number is for the 1st band; *Spot On 6* recycles 22 four-word bundles 34 times in total for this band. Both the numbers of types and tokens for 2k band go down compared to those for 1k band. Except for a few textbooks, those numbers are lower in 2k band even when compared to the ones for 3k band. Interestingly, one textbook, *English 4*, do not include any four-word bundles from the 2nd 1000 list. For the the 3rd 1000, while the number of types is the highest (13) in two textbooks, *Unique 6* and *Texture English 7*, *Texture 7* repeats more four-word bundles, which is 28 in total for this band.

Figure 4.2 visualises the percentages of the four-word bundles covered from the whole reference list. Over the first 3000 four-word bundles in the OANC-spoken, compared to the other textbooks *Unique 6* covers not only many of the bundles (types=43), it also recycles them well (tokens=70) the most among all the textbooks. In *Touchstone 1*, 20 types are included, and they are repeated 33 times in total. Similar to the word-level inquiry results, when all the numbers given in Table 4.4 are overviewed, it can be stated that there is no gradual increase in the number of four-word bundles covered in higher band levels when moved to higher grade textbooks.

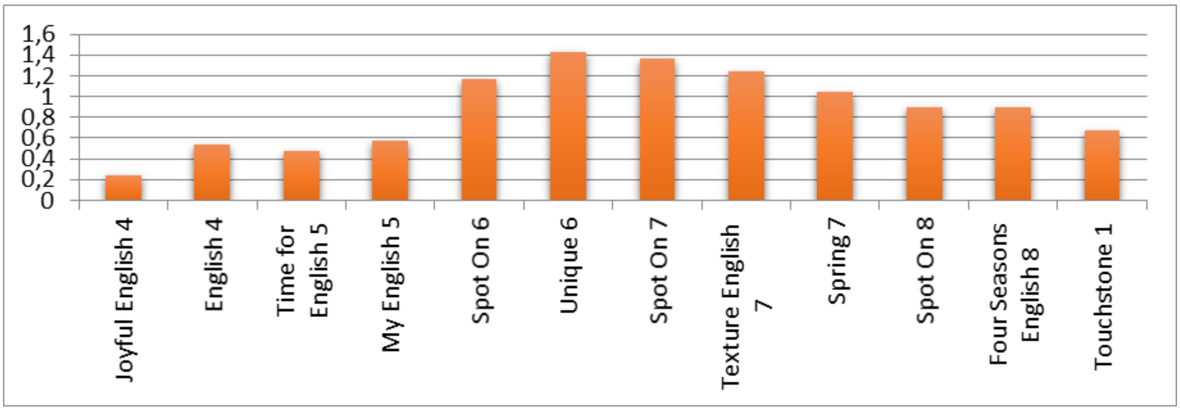


Figure 4.2. The percentages of the four-word bundles covered from the 1k, 2k and 3k reference lists

Table 4.4

Numbers of four-word bundle types and tokens in the TC-Tr and the TC-Ts from 1k, 2k and 3k bands separately and in combination.

Textbook	1k		2k		3k		Total	
	Type	Token	Type	Token	Type	Token	Type	Token
1. Joyful English 4	4	19	0	0	3	4	7	23
2. English 4	7	27	4	9	5	8	16	44
3. Time for English 5	7	17	3	7	4	4	14	28
4. My English 5	9	29	4	9	4	4	17	43
5. Spot On 6	22	34	4	4	9	12	35	50
6. Unique 6	17	25	13	17	13	28	43	70
7. Spot On 7	24	33	7	10	10	17	41	60
8. Texture English 7	17	29	9	12	13	26	39	67
9. Spring 7	14	25	8	11	9	16	31	52
10. Spot On 8	15	17	7	8	5	8	27	33
11. Four Seasons English 8	17	21	4	4	6	6	27	31
12. Touchstone 1	7	15	8	11	5	7	20	33

Table 4.5 represents what percentage of the four-word sequences in the textbooks are the four-word bundles that are in the 1k, 2k and 3k band lists. Considering the total number of words it has (2870), in *Spot On 6* 1.18% of all the words occurred as part of four-word bundles from the 1k band, which is the highest in all the textbooks. For the 2nd and 3rd bands, *Unique 6* has the highest percentage, 0.59 and 0.97, respectively. When the three bands are considered together, although the percentage is very low, the highest percentage is again in *Unique 6*, which means 2.44% of all words were part of the four-word bundles from the reference list.

Table 4.5

Percentages of the four-word bundle bands covered in all the four-word sequences in the TC-Tr and TC-Ts.

Textbook	1k	2k	3k	Total
1. Joyful English 4	0.66	0	0.14	0.8
2. English 4	0.94	0.31	0.28	1.53
3. Time for English 5	0.59	0.24	0.14	0.98
4. My English 5	1.01	0.31	0.14	1.5
5. Spot On 6	1.18	0.14	0.41	1.74
6. Unique 6	0.87	0.59	0.97	2.44
7. Spot On 7	1.15	0.34	0.59	2.1
8. Texture English 7	1.01	0.42	0.91	2.34
9. Spring 7	0.87	0.38	0.56	1.81
10. Spot On 8	0.59	0.28	0.28	1.15
11. Four Seasons English 8	0.73	0.14	0.21	1.08
12. Touchstone 1	0.52	0.38	0.24	1.15

4.1.2. Reference four-word bundles that were never used in the textbooks

Having seen that few four-word bundles occur in the textbook corpora, several four-word bundles from the most frequent 1000 list , which were expected to be encountered with in the textbooks, were checked randomly to see if they were present. Among them, “it s kind of”, “and things like that”, “I mean it s”, “and uh you know” and “yeah that s true” were found to be absent in the textbook corpora. On the other hand, “I don t know” exists in *Spot On 6*, *Spot On 7*, *Spot On 8* and *Touchstone 1*; “I don t think” exists only in *Four Seasons English 8*; “a lot of people” exists only in *My English 5*, “I m not sure” exists only in *Spot On 7*, “what do you think” exists in *Unique 6*, *Texture English 7*, *Spot On 8* and *Four Seasons English 8*; and finally “well I don t” exists in *Spot On 6* and *Texture English 7*.

V. DISCUSSION AND CONCLUSION

5.1. Discussion

In this chapter, the results of the analysis of the textbook corpora will be discussed. The first textbook corpus includes the ELT textbooks approved by the Turkish Ministry of National Education (MoNE) for the 4th, 5th, 6th, 7th and 8th grades (the TC-Tr), and the second one is the corpus of *Touchstone 1* (the TC-Ts), a corpus-informed textbook. The present study uses the term *textbook* to refer to a course book, i.e. student's book (SB), a workbook (WB) and a teacher's book (TB), and both corpora include the listening and conversational parts only, excluding the rest of the written text. While examining the results, it should also be taken into account that the corpora do not include the textbooks as a whole; they are consisted of equal number of words from each SB, WB and TB to enable a comparison between homogeneous samples. For the sake of clarity, the results will be discussed concerning two levels of the study; single word level and four-word bundles level, as discussed in the Results section.

Gavioli and Aston (2001) suggest that corpus helps us make better informed decisions in the syllabus and materials design process. In the same vein, it can also help us evaluate how much better informed the syllabus and materials we use are. According to Schmitt and Schmitt (2012), English language programs should emphasize the teaching of frequently used vocabulary up to 3000 level. As one of the primary sources of language learning, textbooks are expected to present the high-frequency vocabulary so that learners are exposed to linguistic items probable in real life.

Although the results of the present study indicate that most of the single words in the TC-Tr and the TC-Ts come from the most frequent 3000 words in the OANC-spoken, there are some important issues that should be highlighted. It can be inferred from Table 4.1 that *Touchstone 1* seems to aim at teaching the most frequent 1000 words at first hand, compared to its counterparts in the TC-Tr (*Joyful English 4*, *English 4*, *Time for English 5*, and *My English 5*) as it includes both more types and more tokens. When the recycled number of words (tokens) from the top 3000 words in the textbooks is viewed in the Table 4.1, one can also see that although *Touchstone 1* contains fewer types (360) than *My English 5* (384), as one of its counterparts, it contains more of those types (2809). This means that *Touchstone 1* repeats the types of words from the 1k, 2k and 3k reference bands more. Moreover, as seen in Table 4, it is clear that 97.9% of the TC-Ts are made up of the top 3000 words from the OANC-spoken, which makes it more authentic than all the textbooks in the TC-Tr at word level. Obviously, Touchstone 1 is the outstanding one in recycling the words covered from the reference list. On the other hand, among the textbooks of Grade 6, 7 and 8, although *Texture English 7* seems to cover the most frequent items best at single word level, *Spring 7* includes the highest number of word types from the reference band lists as 95.2% of it is from the most frequent 3000 words.

At the 1k band, *Texture English 7* has the highest number of types from the reference list (451 types), which means that it includes 45.1% of the most frequent 1000 words on the list based on the OANC-spoken (Table 4.3). This indicates that none of the textbooks approved by the MoNE represents even the half of the spoken American English regarding its most frequently used 1000 words. As for the 2k and 3k bands, there is a sharp decrease in the percentages covered by the both corpora. Among all the textbooks, *Spot On 6* covers the words best both from the 2k and 3k bands, with 11.3% and 6.4%, respectively. On the other hand, when all the three bands are considered, *Texture English 7* covers the most types; 59.1% of the words from the reference list. As Table 4.3 shows, considering the most frequent 3000 words in the OANC-spoken, only three textbooks approved by the MoNE (*Spot On 6*, *Texture 7* and *Spot On 8*) mirror slightly more than half of all the words from the reference list.

The table Nation offers (2006, p. 79; see Figure 3) indicates that vocabulary (as word families) needed to understand a text is mostly from the 1k band. Then, the words from 2k and 3k bands are less required, though helps one become more productive in written or spoken language. Yet, when looked at the general picture regarding the three bands, students in Turkey are exposed to only slightly more than the half of the most frequently used words in spoken English by means of the textbooks approved by the MoNE at the end of a five-year-education of English language. When the fact that this study considers word in non-lemmatized forms, not in word families, is taken into account, it can be implied that the number of the word families introduced to students via textbooks is very few. As for the vocabulary content of *Touchstone 1*, one can see that at 1k band it has the highest number of types among its counterparts. However, despite the fact that *Touchstone 1 is* a corpus-informed textbook, it is not the one that covers the highest number of types from the whole list of 3000 most frequent words (360 types).

Regarding the four-word bundles, the study reveals that both of the textbook corpora include a small number of bundles. Out of 3000 most frequent four-word bundles, 43 types of them are included in *Unique 6*, and they are recycled 70 times, the highest among all the textbooks. Yet, at the 1k band, *Spot On 7* is the one covering the highest number of four-word bundle types. Surprisingly again, *Touchstone 1* does not seem to outperform its counterparts in the 1k and 2k searches, although the number of four-word bundles it covers is more than its counterparts when all the bands are considered in total.

As Table 4.5 indicates, both of the textbook corpora are poor in presenting students the frequently used four-word bundles for spoken language while sequencing the words. This means that learners using the textbooks employed in this study are exposed to few of the bundles in the reference list. As textbooks are primary sources of the target language for many students, the amount of bundles they are exposed to is crucial in language learning process.

Results indicate that different textbooks designed for a particular grade are different from each other in terms of their single word and four-word bundle content for spoken language. Similarly, in line with Ljung's (1999) findings, it can also be drawn from the results that there is a flattered profile through different grades. Although one might expect the difficulty of the words grows successively, with less occurrences from the 1k band list and more from the 3k band list when moving up through the grades, even in a series prepared by the same author, *Spot On 6-7-8*, that cannot be observed. According to Milton (2009), there is little agreement on the criteria for textbook vocabulary selection, which is also the case for the textbooks approved by the Turkish MoNE. Their authors do not clearly state the criteria they rely on. Although the authors of the textbooks in TC-Tr argue in the forewords of the teachers' books that they observed the importance of the real life while preparing the content, the present evaluation of the textbooks through corpus analysis indicates that they ignored the frequency information of the actual spoken language as a criterion considerably. While designing textbooks Schmitt and Schmitt (2012) argue that the most frequent 3000 word families are needed "to largely understand (and presumably produce) conversational English" (p. 7). With the emphasis on communication in language teaching (Nation & Waring, 1997), learners need to be exposed to more occurrences of real language use. Consequently, it can be suggested that the single word and bundle content of the textbooks be improved.

5.2. Pedagogical Implications

Considering the fact that the textbooks approved by the Turkish MoNE are used by a great number of students in state schools in Turkey, the pedagogical implications that can be drawn from the study are of valuable importance for three different groups; educational authorities, language textbook evaluators and designers, and teachers. To begin with, considering the fact that, depending on the textbooks used, students are exposed to different number of words and bundles with different frequencies, which may result in differences in language learning. This indicates that there is a lack of predefined systematic criteria for the textbooks regarding their number and type of the words included. As mentioned in the methodology section, even the target level of language learning is not indicated in the EFL textbooks, let alone their vocabulary content. Under these circumstances, it is not surprising to have different numbers of word types and tokens covered from the three bands. For each grade, a target level should be defined based on an accepted standardization method such as Common European Framework (CEF), and the target vocabulary that students are aimed to be taught at the end of the education should be specified.

Ljung (1991) argues that by merely looking at its frequency list of words one can obtain a significant amount of information about a text. Taking this into account, utilizing corpus in textbook evaluation process can provide us valuable information about what it offers and how suitable it is for our purposes in language teaching/learning.

From the discrepancies between corpus-driven data and what the textbooks include, one can infer that to achieve a higher degree of authenticity in language learning, corpus is a useful method that can be employed in EFL textbook production, improvement and evaluation. Indeed, corpus renders language textbook authors’ job an easier task since they can base their production on empirical findings, rather than their intuitions. By making more reference to corpus findings and using the frequency and context information of the language already spoken/written, textbook evaluators and authors, and can increase the possibility of meaningful input to be provided to learners. Therefore, corpus should be consulted while considering what to include in ELT textbooks and what to exclude from them.

Moreover, textbook authors should reflect any aspect of spoken language, including repetitions and fillers such as *uh*, *er*, *um*, in the transcriptions as much as possible both in student course book and in teacher’s book because their occurrence can be highlighted more efficiently in this way. The difference between the written and spoken language should be indicated so that learners could gain the ability to use different codes appropriately. Taking for granted that materials developers are interested in frequency information (Biber & Reppen, 2002), a list of common words and bundles in spoken language can be provided for learners in textbooks.

On the other hand, the present study does not suggest that vocabulary and bundles existing in spoken corpus data is the only criteria that should be taken into account while designing and evaluating ELT textbooks. Other factors, such as the frequency information of grammar structures can also be considered helpful. Still, a closer look at the vocabulary in real spoken language use through language corpora and comparing the content in textbooks with them would shed light on the authenticity of the language offered in textbooks, and the efforts to teach vocabulary that is observed in use might increase the possibility for learners to communicate successfully with competent speakers of English.

As for teachers, it seems necessary for them to detect and introduce the high frequency vocabulary to the learners in order to compensate for the inadequacy of the textbook content. They can achieve this through extra language materials by adapting authentic written/spoken texts as well as being a target model for the learners. Especially for spoken language, it is the teachers that can demonstrate many performance phenomena such as repetitions and filled pauses (Mukherjee, 2009, p. 225). Additionally, syllabus design can be improved significantly if they can focus on the words that cover the majority of written and spoken productions in their teaching.

5.3. Limitations of the Study

Basically, there are four limitations of the study. First, as a reference corpus, the OANC-spoken was employed, which represents only the American spoken language. However, in the TC-Tr it is not clear if there is one variety of English that the textbooks stick to. If there is one and it is not American English, using an American corpus to compare the TC-Tr with might have led to fewer occurrences from the bands in those textbooks. because of two reasons; first, the differences in the transcriptions of the words in American and British English, such as *behavior* and *behaviour*; and second, proper nouns occurring specifically in the context of the variety, such as *Virginia*, *Florida* and *Washington* in the most frequent 3000 words list. Therefore, it might have been a better idea to use a reference corpus representing more than one variety of English.

Secondly, it was not possible to observe the developmental change regarding the coverage of the reference bands through the *Touchstone* series as the level of English increases. It was indicated in the results that *Touchstone 1* covers words from the 1k band the best compared to its counterparts, but not in the 2k and 3k bands. The reason might be because its authors believe that the higher level vocabulary should be introduced in the higher levels of the series. However, as the study included only the first level of the *Touchstone* series, it is not possible to have a clear answer.

Another limitation was the fact that the pedagogical corpora was formed by including a definite amount of words from the beginning of the textbook. The language used might differ in its level of lexical choice in the middle or at the end of a textbook, which might lead to different results.

Finally, some features of spoken language such as repetitions and filled pauses cannot be expected to be reflected perfectly in textbooks. To illustrate, a repetition can be expressed by repeating the word(s) twice in written text although it is repeated for, say, three times in a conversation listened to by students. A filled pause can be indicated in transcription with different spellings. While in British English *er* and *erm* are preferred, in American English filled pauses are indicated as *uh* and *um* (Department of Linguistics, n.d.). Furthermore, the way the textbooks spelt them were not standardized, and varied among textbooks. Although they used a similar spelling as given in British or American English, they were observed to duplicate the last sound, like *errrr* or *uhh*. Yet, the transcriptions were taken as provided in the textbooks. Consequently, some features of spoken English in the bands could not been found in the textbooks.

5.4. Suggestions for Further Research

As a limitation of the study, the reference corpus choice has already been indicated. To eliminate this limitation, a corpus representing more varieties of English can be included in a similar study. Moreover, the textbooks can also be compared with a general corpus representing another variety of English such as British National Corpus (BNC) to explore if the study would yield different results.

It is doubtful whether students would have rich vocabulary and bundle knowledge at the end of the secondary school in Turkey. A study investigating the relationship between the vocabulary content in the textbooks as a primary language learning source and student spoken and/or written productions can be conducted. Such a study might point out the effect of the frequency of vocabulary and bundle occurrences in the textbooks, as well. Moreover, in that study the TC-Tr can be expanded in future research studies in two ways;

by including the non-conversational parts of the textbooks, and the textbooks approved by the Ministry of National Education (MoNE) to be studied in schools other than secondary schools, such as high schools.

In the present study, a corpus-informed textbook *Touchstone 1* was examined. Although the levels of most of the textbooks in the TC-Tr are not clearly stated, one can assume that the textbooks for Grades 4 and 5 match the best with *Touchstone 1*. Having compared it with its counterparts in the TC-Tr, the reason why it is not the outstanding one might be because it is the first book of the series. *Touchstone* series are stated to address to A1 to B2 learners. That is why to have an idea about the benefits of corpora in evaluating textbooks it might be a better approach to study the whole series.

In addition, to enable a comparison between a corpus-informed textbook corpus and the TC-Tr at all levels, other corpus-informed textbooks such as *Real grammar* (Conrad, & Biber, 2009), and *Grammar and beyond* (Reppen, Bunting, Diniz, Blass, Iannuzzi, & Savage, 2012) can be included in further studies. Having seen that in the TC-Tr there is an inconsistent increase in the numbers of types and tokens for four-word bundles of higher band levels as the grade goes up, a study including all the *Touchstone* textbooks can also discover the tendency between the bands and volumes.

Moreover, the same study can be carried out by using the lemmatized forms of vocabulary to determine the most frequently used 3000 single words and four-word bundles in the reference corpus. The knowledge of word families, rather than lemmas, are also considered to be significant indicators of language competence. Nation (2006), for example, notes that to comprehend 98% of a text, vocabulary of 8,000 to 9,000 word families for written text and vocabulary of 6,000 to 7,000 for spoken text is needed. Considering this, a study discovering to what extent the textbooks cover the most frequently used word families can be conducted.

5.5. Conclusion

Taking the few number of single words from the reference bands into account, the study has clearly demonstrated that most of the textbooks approved by the MoNE represent authentic spoken English in a very limited way. One can expect to encounter much fewer word families among the most frequently used 3000, which Schmitt and Schmitt (2012) suggest to be crucial to understand - and presumably produce - conversational English. Even if all the textbooks in the TC-Tr are supposed to be studied and their lexical content is learned perfectly —which is impossible in practice as there are more than one textbook for a grade, the lexical content of the TC-Tr textbooks is far away from offering the most frequently used 3000 word families so that the learners learn them in order to largely understand and produce conversational English. On the other hand, when *Touchstone 1* is considered, although it is in a better condition compared to its counterparts, it still lacks a great number of frequently used single words and four-word bundles occurring in authentic spoken English.

The study suggests that corpus-informed textbooks should be encouraged, and the ones available in the market be revised so that they can better reflect the vocabulary and formulaic language in authentic English. EFL textbooks are of great importance to many students at state schools since they are the only language source of spoken language available to them. That is why while preparing and evaluating EFL textbooks, authors and publishers should bear in mind that the target words and bundles are in dialogues and other parts that reflect the spoken language.

REFERENCES

Altenberg, B. & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*, 173-94.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D., & Reppen, R. (2002).What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, *24*, 199-208.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263-286.

Biber, D. & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard, & S. Oksefjell, (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp.181-189). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at …:* Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Cambridge English CEFR Correlations. Retrieved February 24, 2014 from http://www.cambridge.org/servlet/file/TouchstoneCEFR2012_Level1.pdf?ITEM_ENT_ID =7176261&ITEM_VERSION=2&COLLSPEC_ENT_ID=7.

Carter, R.A. (1998). Orders of reality: CANCODE, communication and culture. *ELT Journal 52*(1), 43-56.

Chen, L. (2010). An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. In D. Wood (Ed.) *Perspectives on formulaic language: acquisition communication* (pp. 107-125). London; New York: Continuum.

Chen, Y.H. & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*, 30-49.

Conrad, S. & Biber, D. (2009). *Real grammar: A corpus-based approach to instruction*. New York: Pearson Longman.

Cook, V. (2001). The philosopher pulled the lower jaw of the hen. Ludicrous invented sentences in language teaching. *Applied Linguistics, 22(3),* 366-387.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17*(4), 391-406.

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora, 3*, 4-57.

Cowie, A. P. (1992). Multi word lexical units and communicative language teaching. In Arnaud, P. J. L., & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp.1-12). London: Macmillan.

Criado, R., & Sánchez, A. (2009). Vocabulary in EFL Textbooks. A Contrastive Analysis against Three Corpus-Based Word Ranges. In A. Sánchez & P. Cantos (Eds.), *A Survey on corpus-based research / Panorama de investigaciones basadas en corpus* (pp. 862-875). Murcia: Editum (Servicio de Publicacio nes de la Universidad de Murcia).

Cullen, R., & Kuo, I. V. (2007). Spoken grammar and ELT course materials: A missing link? *TESOL Quarterly,* 41, 361-386.

Davidson, D. J., Indefrey, P., & Gullberg, M. (2008). Words that second language learners are likely to hear, read, and use. *Bilingualism: Language and Cognition, 11*(1), 133-146.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*(1), 59-80.

Dubin, F. (1995). The craft of materials writing. In P. Byrd (Ed.), *Materials writer's guide* (pp. 13-22). Boston: Heinle & Heinle.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163-188.

Flowerdew, L. 2012. *Corpora and Language Education*. Basingstoke: Palgrave Macmillan.

Gardner, D. (2007). Validating the construct of *word* in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241-265.

Gavioli, L. & Aston, G. (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal*, *55*(3), 238-246.

Goethals, M. (2003). E.E.T.: the European English Teaching vocabulary-list. In B. Lewandowska-Tomaszczyk (Ed.). *Practical applications in language and computers* (pp. 417-427). Frankfurt: Peter Lang.

Gray, J. (2002). The global course book in English language teaching. In D. Block and D. Cameron (Eds.), *Globalization and language teaching* (pp. 151–67). London: Routledge.

Greenbaum, S. (1990). Standard English and the international corpus of English. *World Englishes, 9*, 79-83.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, *24*, 287-298.

Harwood, N. (2005). What do we want EAP teaching materials for?. *Journal of English for Academic Purposes, 4*(2), 149-161.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32,* 150-169.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Jalali H, & Ghayoomi S. (2010). A comparative qualitative study of lexical bundles in three academic genres of applied linguistics. *MJAL, 2*(4), 323-333.

Jablonkai, R. (2009). In the light of: A corpus-based analysis of lexical bundles in two EU related registers. *WoPaLP, 3,* 1-27. Retrieved February 24, 2014 from http://langped.elte.hu/WoPaLParticles/W3Jablonkai.pdf

Kim, Y. (2009). Korean lexical bundles in conversations and academic texts. *Corpora, 4(2),* 135-165.

Klages, M., & Römer, U. (2002). Translating Modal Meanings in the EFL Classroom. In S. Scholz, M. Klages, E. Hantson & U. Römer (Eds.), *Language: Context and cognition papers in honour of Wolf-Dietrich Bald's 60th birthday* (pp. 201-216). Munich: Langenscheidt-Longman.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman. 8-29.

Leech, G. (1997). Teaching and language corpora: a convergence. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.) *Teaching and language corpora* (pp. 1-23). London: Longman.

Madden, C. G., & T. N. Rohlck. (2000). *Discussion & interaction in the academic community*. Ann Arbor: University of Michigan Press.

Martini, O. P. J. (2012). *High Frequency Vocabulary in a Secondary Quebec ESL Textbook Corpus*. Master's thesis, Concordia University. Retrieved February 24, 2014 from http://spectrum.library.concordia.ca/974698/.

Matsuda, A., & Friedrich, P. (2012). Selecting an instructional variety for an EIL curriculum. In A. Matsuda (Eds), *Principles and practices of teaching English as an international language*, (pp.17-27). Bristol: Multilingual Matters.

Matsuoka, W., and Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language, 22*(1), 56-70.

McCarthy, M. (2004). *Touchstone: from corpus to course book.* Cambridge: Cambridge University Press.

McCarthy, M. & Carter R.A. (2002). This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Teanga, 21*, 30-52.

McDonough, J. & Shaw, C. (2003). *Materials and methods in ELT: A teacher's guide*. Oxford: Blackwell.

McEnery, T. & Wilson, A. (2001). *Corpus linguistics.* Edinburgh: Edinburgh University Press.

McEnery, T. & Xiao, R. (2010). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 364-380). London and New York: Routledge.

Meunier F., & Gouverneur C. (2007). The treatment of phraseology in ELT textbooks. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom: Selected papers from the sixth international conference on Teaching and Language Corpora (TaLC 6), University of Granada,* Spain, 4–7 July, 2004: Volume 61 (pp. 119–139). Amsterdam: Rodopi.

Meunier, F. and Gouverneur, C. (2009). New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (Ed.). *Corpora and language teaching*. Amsterdam & Philadelphia: Benjamins.

Milton, J. (2009). Vocabulary acquisition and classroom input. In J. Milton *Measuring second language vocabulary acquisition*. (pp. 193-217). Bristol: Multilingual Matters.

Mindt, D. (1996). English corpus linguistics and the foreign language teaching syllabus. In J. Thomas & M. Short (Eds.), *Using corpora for language research*, (pp. 232-247). London: Longman.

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (Ed.) *Corpora and language teaching*, (pp. 203-330). Amsterdam: John Benjamins.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59-82.

Nation, P. and Waring, R. (1997). Vocabulary size, text coverage, and word lists. In Norbert Schmitt & Michael McCarthy (Eds.). *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge, Cambridge University Press.

Nattinger, J. R. and DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. **(**2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

O’Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching.* Amsterdam: John Benjamins.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

Reppen, R., Bunting, J., Diniz, L., Blass, L., Iannuzzi, S. & Savage, A. (2012). *Grammar and beyond,* (Vols. 1–4), Cambridge: Cambridge University Press.

Römer, U. (2004a). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini & D. Stewart (Eds.), *Corpora and language learners* (pp.151-168). Amsterdam: John Benjamins.

Römer, U. (2004b). A Corpus-driven approach to modal auxiliaries and their didactics. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp.185-199). Philadelphia, Pa.: John Benjamins.

Römer, U. (2005a). Looking at looking: Functions and contexts of progressives in spoken English and 'school' English. In A. Renouf & A. Kehoe (Eds.), *The changing face of corpus linguistics* (pp. 231-242). Amsterdam: Rodopi.

Römer, U. (2005b). *Progressives, patterns, pedagogy: A Corpus-driven approach to English progressive forms, functions, contexts and didactics.* Amsterdam: John Benjamins.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31,* 205-225.

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*, 72-78.

Shirato, J. & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research, 11*(4), 393-412.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Hampshire: Palgrave and Macmillan.

Schmitt, N. (2012). Formulaic language. In C. A. Chapelle (Ed.), *The encyclopaedia of applied linguistics.* Wiley-Blackwell.

Schmitt, N. & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*.

Stubbs, M. (2004). Language corpora. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 106-132). Oxford: Blackwell.

Super, A. K. (2004). *A corpus-based approach to ESL: Textbook and materials development and evaluation*. Master's thesis, Michigan State University. Retrieved February 24, 2014 from http://www.editlib.org/p/126829.

Thornbury, S. (2002). *How to teach vocabulary*. Harlow: Longman.

Widdowson, H., G. (2000). On the limitations of linguistics applied. *Applied Linguistics, 21*(1), 3-25.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*, 463-489.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *57*(4), 541-572.

Corpora

British National Corpus (BNC). Retrieved from http://corpus.byu.edu/bnc/.

Corpus of Contemporary American English (COCA). Retrieved from http://corpus.byu.edu/coca/.

Open American National Corpus (OANC). Retrieved from http://www.anc.org/data/oanc/download/.

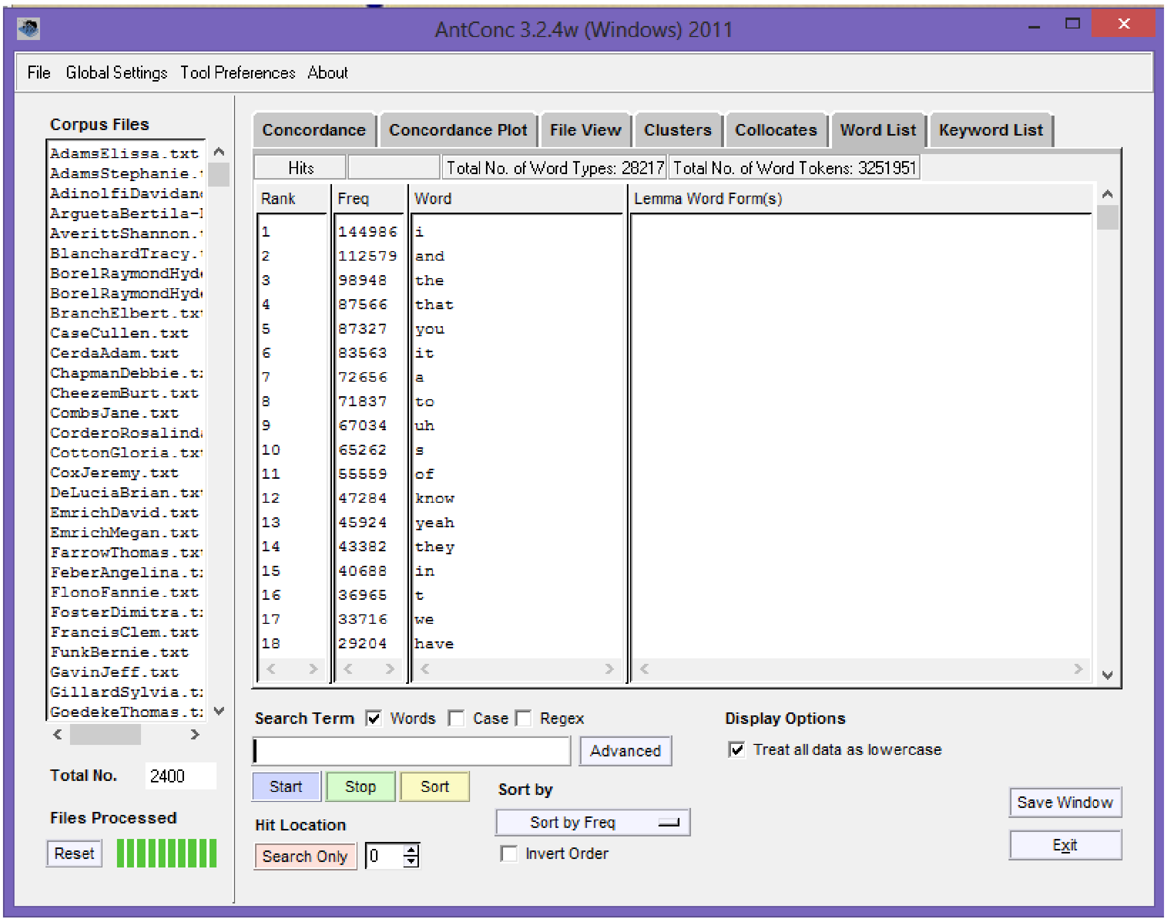
Concordancer

Anthony, L. (2011). AntConc (Version 3.2.4w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

APPENDICES

APPENDIX A1

Screenshot showing the analysis of the most frequent 3000 words in the OANC-spoken



Corpus Files

AdamsElissa.txt
AdamsStephanie.t
AdinolfiDavidan
ArguetaBertila-I
AverittShannon.t
BlanchardTracy.t
BorelRaymondHyd
BorelRaymondHyd
BranchElbert.tx
CaseCullen.txt
CerdaAdam.txt
ChapmanDebbie.t:
CheezemBurt.txt
CombsJane.txt
CorderoRosalind
CottonGloria.tx
CoxJeremy.txt
DeLuciaBrian.tx
EmrichDavid.txt
EmrichMegan.txt
FarrowThomas.tx
FeberAngelina.t:
FlonoFannie.txt
FosterDimitra.t:
FrancisClem.txt
FunkBernie.txt
GavinJeff.txt
GillardSylvia.t:
GoedekeThomas.t:

Total No. 2400

Files Processed

Hits | Total No. of Word Types: 28217 | Total No. of Word Tokens: 3251951

Rank	Freq	Word	Lemma Word Form(s)
1	144986	i	
2	112579	and	
3	98948	the	
4	87566	that	
5	87327	you	
6	83563	it	
7	72656	a	
8	71837	to	
9	67034	uh	
10	65262	s	
11	55559	of	
12	47284	know	
13	45924	yeah	
14	43382	they	
15	40688	in	
16	36965	t	
17	33716	we	
18	29204	have	

Search Term ☑ Words ☐ Case ☐ Regex

Display Options

☑ Treat all data as lowercase

Sort by

Sort by Freq

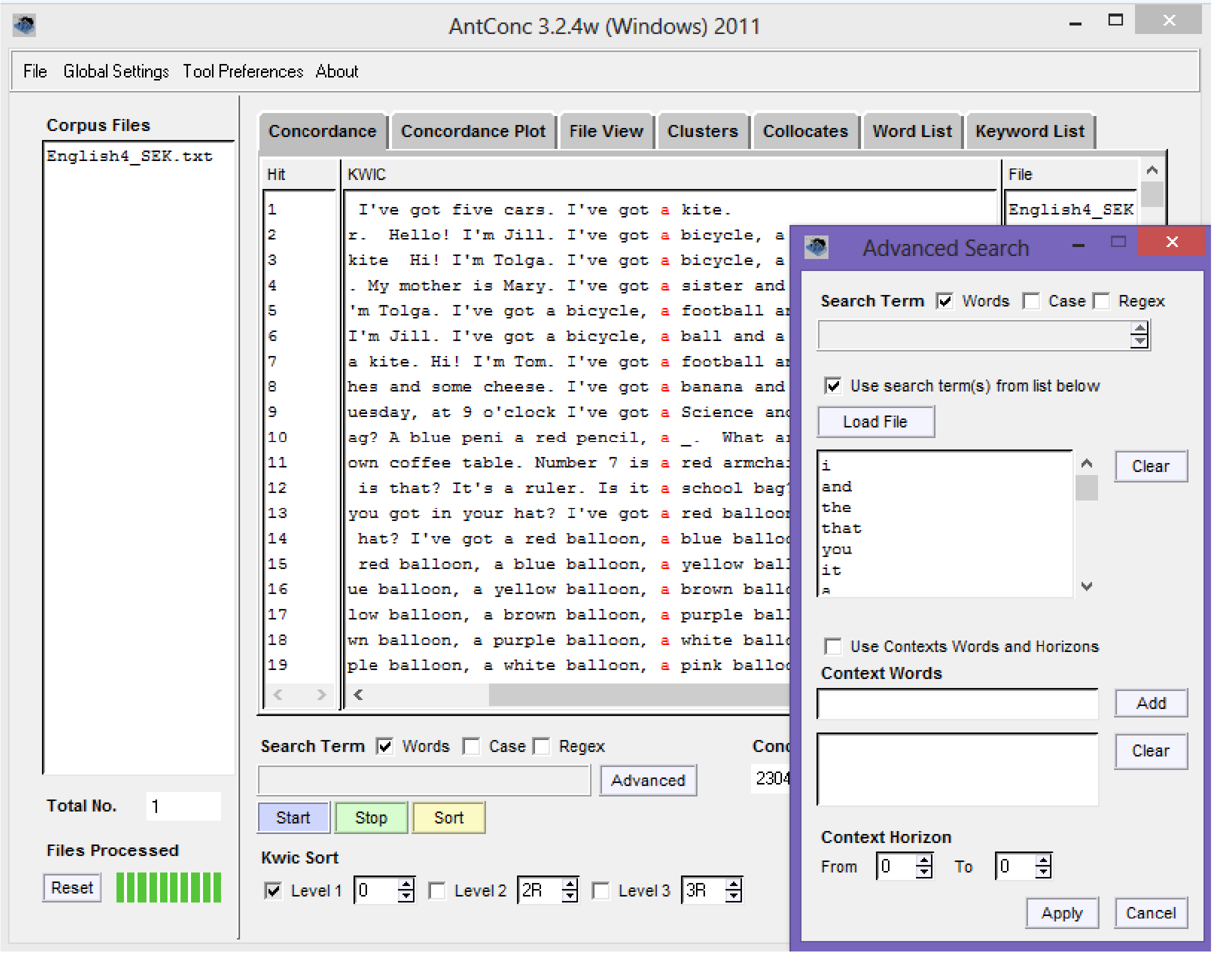
☐ Invert Order

Hit Location

Search Only 0

APPENDIX A2

Screenshot showing the advanced search in a textbook for the most frequent 1000 words retrieved from OANC-spoken



Corpus Files
English4_SEK.txt

Total No. 1

Files Processed

```
Hit  KWIC                                                File
1     I've got five cars. I've got a kite.               English4_SEK
2    r.  Hello! I'm Jill. I've got a bicycle, a
3    kite  Hi! I'm Tolga. I've got a bicycle, a
4    . My mother is Mary. I've got a sister and
5    'm Tolga. I've got a bicycle, a football ar
6    I'm Jill. I've got a bicycle, a ball and a
7    a kite. Hi! I'm Tom. I've got a football ar
8    hes and some cheese. I've got a banana and
9    uesday, at 9 o'clock I've got a Science and
10   ag? A blue peni a red pencil, a _.  What ar
11   own coffee table. Number 7 is a red armchai
12    is that? It's a ruler. Is it a school bag
13   you got in your hat? I've got a red balloor
14    hat? I've got a red balloon, a blue balloo
15    red balloon, a blue balloon, a yellow ball
16   ue balloon, a yellow balloon, a brown ballo
17   low balloon, a brown balloon, a purple ball
18   wn balloon, a purple balloon, a white ballo
19   ple balloon, a white balloon, a pink balloo
```

Search Term ☑ Words □ Case □ Regex

Conc 2304

Kwic Sort

☑ Level 1 0 □ Level 2 2R □ Level 3 3R

Advanced Search

Search Term ☑ Words □ Case □ Regex

☑ Use search term(s) from list below

```
i
and
the
that
you
it
a
```

□ Use Contexts Words and Horizons

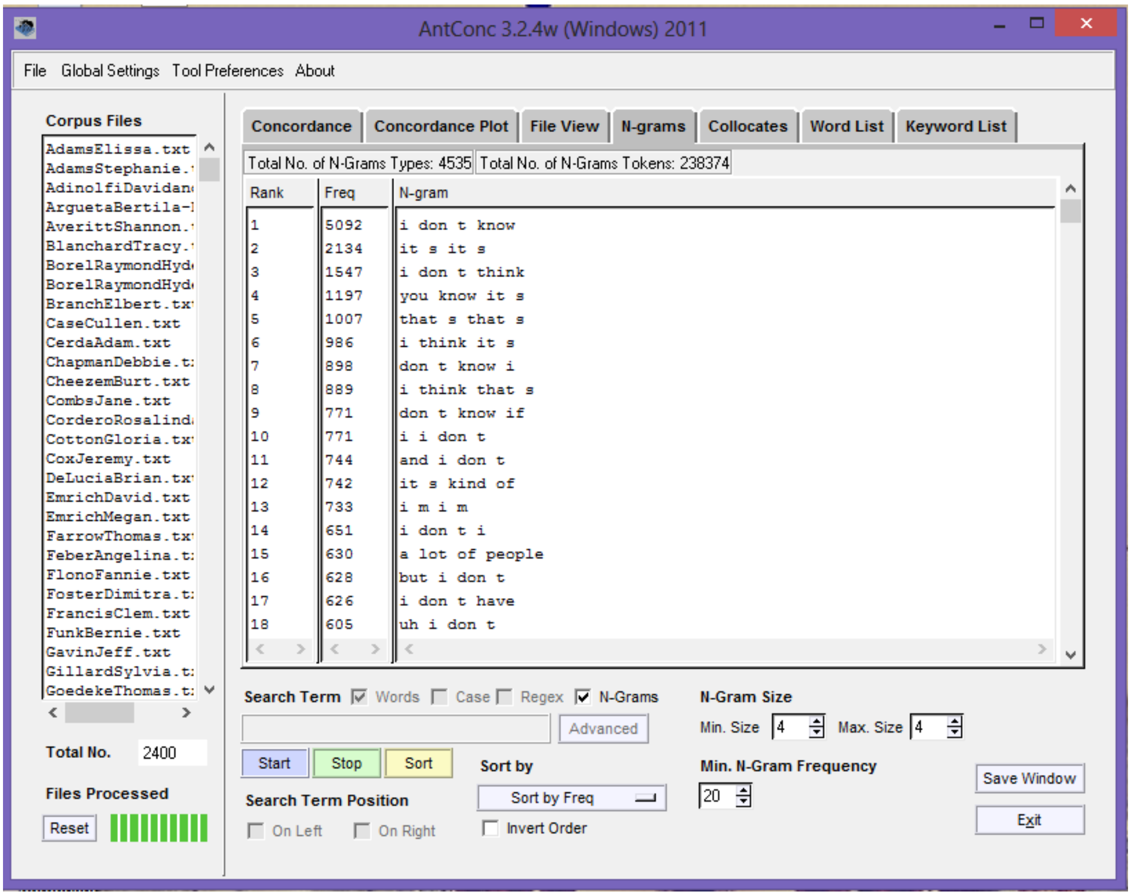
Context Words

Context Horizon

From 0 To 0

APPENDIX B1

Screenshot showing the analysis of the most frequent 3000 four-word bundles in OANC-spoken



Corpus Files

AdamsElissa.txt
AdamsStephanie.t
AdinolfiDavidan
ArguetaBertila-I
AverittShannon.t
BlanchardTracy.t
BorelRaymondHyd
BorelRaymondHyd
BranchElbert.tx
CaseCullen.txt
CerdaAdam.txt
ChapmanDebbie.t:
CheezemBurt.txt
CombsJane.txt
CorderoRosalind
CottonGloria.tx
CoxJeremy.txt
DeLuciaBrian.tx
EmrichDavid.txt
EmrichMegan.txt
FarrowThomas.tx
FeberAngelina.t:
FlonoFannie.txt
FosterDimitra.t:
FrancisClem.txt
FunkBernie.txt
GavinJeff.txt
GillardSylvia.t:
GoedekeThomas.t:

Total No. 2400

Files Processed

Total No. of N-Grams Types: 4535 | Total No. of N-Grams Tokens: 238374

Rank	Freq	N-gram
1	5092	i don t know
2	2134	it s it s
3	1547	i don t think
4	1197	you know it s
5	1007	that s that s
6	986	i think it s
7	898	don t know i
8	889	i think that s
9	771	don t know if
10	771	i i don t
11	744	and i don t
12	742	it s kind of
13	733	i m i m
14	651	i don t i
15	630	a lot of people
16	628	but i don t
17	626	i don t have
18	605	uh i don t

Search Term ☑ Words □ Case □ Regex ☑ N-Grams

N-Gram Size: Min. Size 4 Max. Size 4

Sort by: Sort by Freq

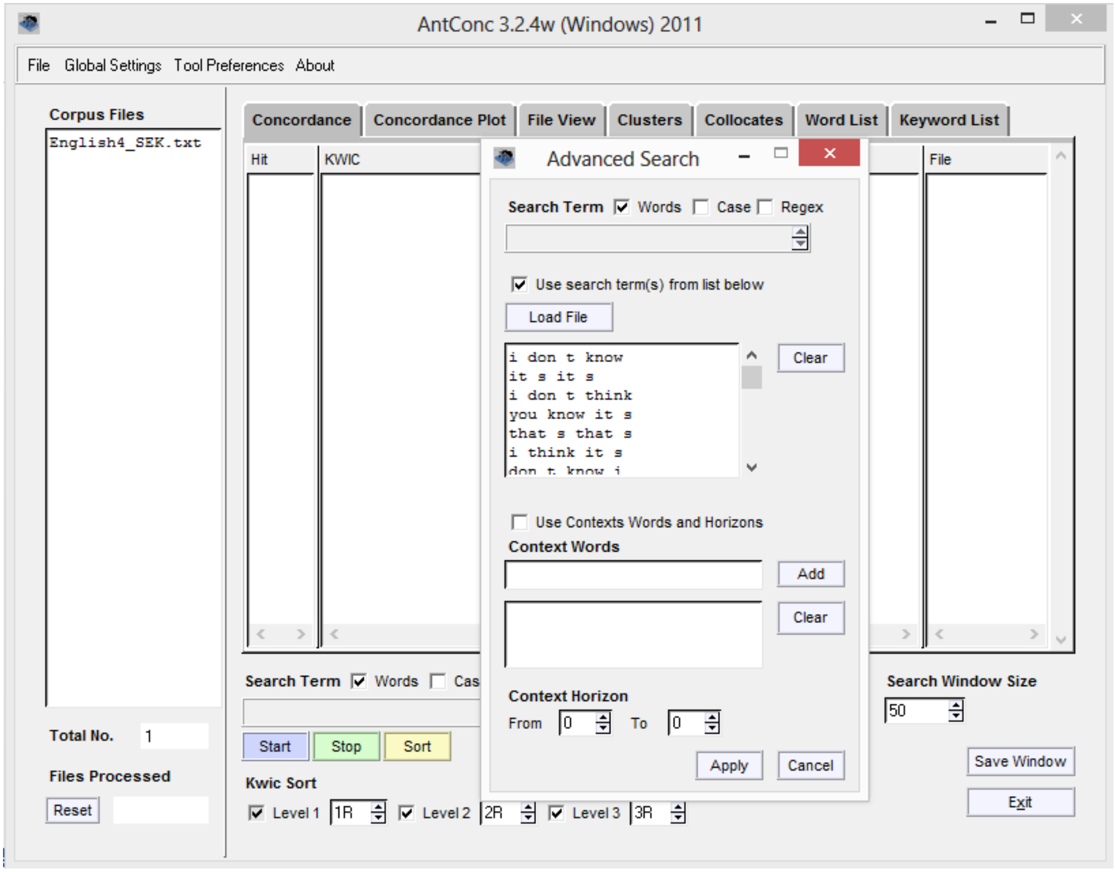
Min. N-Gram Frequency: 20

Search Term Position: □ On Left □ On Right

□ Invert Order

APPENDIX B2

Screenshot showing the advanced search in a textbook for the most frequent 1000 four-word bundles retrieved from OANC-spoken



Corpus Files

English4_SEK.txt

Total No. 1

Search Term ☑ Words □ Case □ Regex

☑ Use search term(s) from list below

```
i don t know
it s it s
i don t think
you know it s
that s that s
i think it s
don t know i
```

□ Use Contexts Words and Horizons

Context Horizon

From 0 To 0

Search Term ☑ Words □ Cas

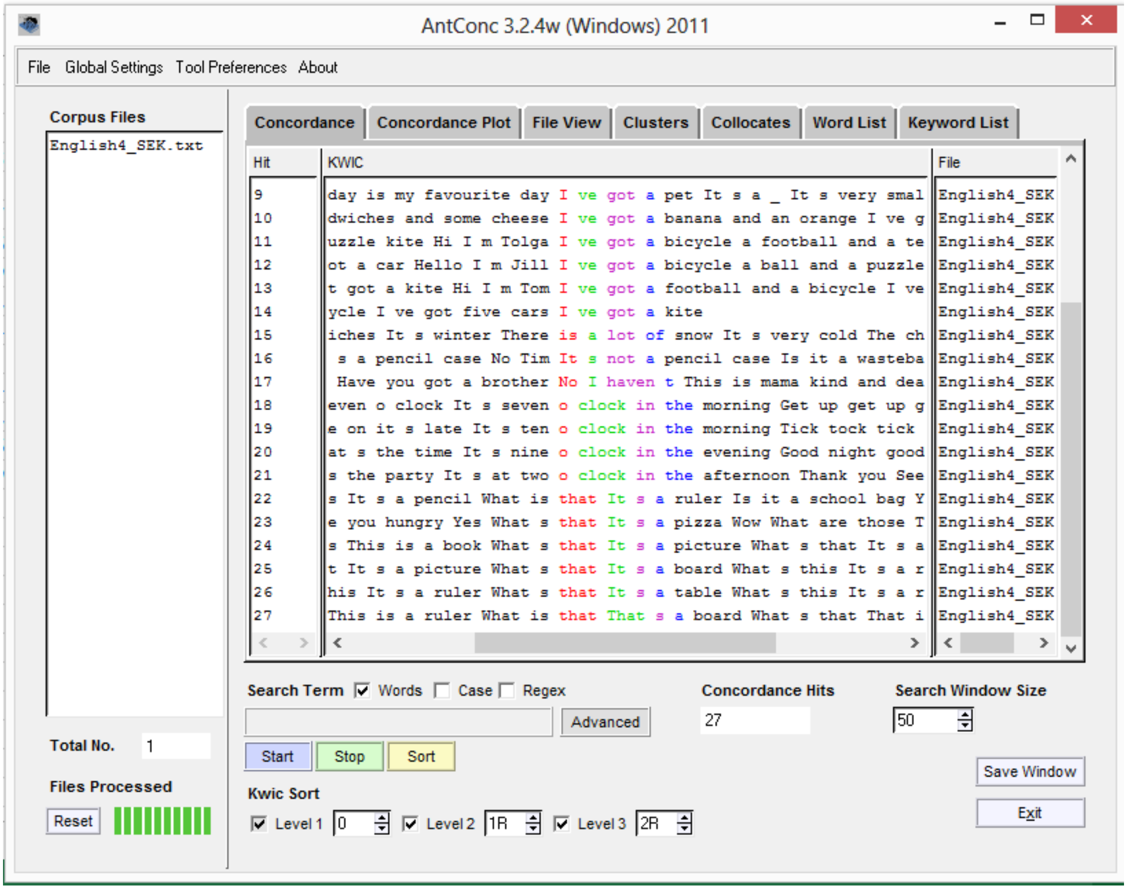
Search Window Size 50

Kwic Sort

☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

APPENDIX C

Screenshot showing the advanced search results in a textbook for the most frequent 1000 four-word bundles retrieved from OANC-s, in alphabetical order.



Corpus Files

English4_SEK.txt

Hit	KWIC	File
9	day is my favourite day I ve got a pet It s a _ It s very smal	English4_SEK
10	dwiches and some cheese I ve got a banana and an orange I ve g	English4_SEK
11	uzzle kite Hi I m Tolga I ve got a bicycle a football and a te	English4_SEK
12	ot a car Hello I m Jill I ve got a bicycle a ball and a puzzle	English4_SEK
13	t got a kite Hi I m Tom I ve got a football and a bicycle I ve	English4_SEK
14	ycle I ve got five cars I ve got a kite	English4_SEK
15	iches It s winter There is a lot of snow It s very cold The ch	English4_SEK
16	s a pencil case No Tim It s not a pencil case Is it a wasteba	English4_SEK
17	Have you got a brother No I haven t This is mama kind and dea	English4_SEK
18	even o clock It s seven o clock in the morning Get up get up g	English4_SEK
19	e on it s late It s ten o clock in the morning Tick tock tick	English4_SEK
20	at s the time It s nine o clock in the evening Good night good	English4_SEK
21	s the party It s at two o clock in the afternoon Thank you See	English4_SEK
22	s It s a pencil What is that It s a ruler Is it a school bag Y	English4_SEK
23	e you hungry Yes What s that It s a pizza Wow What are those T	English4_SEK
24	s This is a book What s that It s a picture What s that It s a	English4_SEK
25	t It s a picture What s that It s a board What s this It s a r	English4_SEK
26	his It s a ruler What s that It s a table What s this It s a r	English4_SEK
27	This is a ruler What is that That s a board What s that That i	English4_SEK

Search Term ☑ Words □ Case □ Regex

Concordance Hits 27

Search Window Size 50

Kwic Sort ☑ Level 1 0 ☑ Level 2 1R ☑ Level 3 2R

Total No. 1

Files Processed