High Frequency Vocabulary in a Secondary Quebec ESL Textbook Corpus

Juliane Oliveira Pisani Martini

Department of Education

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Arts (Applied Linguistics) at

Concordia University

Montreal, Quebec, Canada

August 2012

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Juliane Oliveira Pisani Martini

Entitled: High Frequency Vocabulary in a Secondary Quebec ESL Textbook Corpus

and submitted in partial fulfilment of the requirements for the degree of

Master of Arts (Applied Linguistics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Teresa Hernández González	Chair
Joanna White	Examiner
Sara Kennedy	Examiner
Marlise Horst	Supervisor

Approved by Richard Schmid
Chair of Department

Bryan Lewis
Dean of Faculty

Date August 16, 2012

Abstract

High Frequency Vocabulary in a Secondary Quebec ESL Textbook Corpus

Juliane Oliveira Pisani Martini

It is widely accepted that high-frequency vocabulary must be taught in ESL/EFL classrooms, and that learners benefit from learning it (Schmitt, 2011; Nation, 2001). Research also confirms that recycling vocabulary is beneficial in facilitating the acquisition of vocabulary knowledge Milton (2009). In order to understand the lexical characteristics of classroom input secondary ESL learners in Québec are exposed to, I gathered and analyzed a corpus of Ministry-approved textbooks (*Collection Quest* for cycle 2 published by Chenelière Éducation). Subsequently, I used the corpus findings to develop a pedagogical word list that targets high frequency words that may need more emphasis or be missing entirely. Results showed that there are considerable deficits in the vocabulary presented in the three books. The 1k level was considered well represented because most words at this level occur in the books frequently and are widely recycled across the volumes. At the 2k and 3k levels, most words also occur in the books; however their frequency and range of occurrence are not satisfactory in terms of promoting successful acquisition. As for the mid-frequency vocabulary, results show that students have very few opportunities to encounter these words in their books. Most of the words between the 3k and the 9k levels are not recycled frequently in the corpus.

Acknowledgements

It is difficult to find the right words to fully express my gratitude to my supervisor, Marlise Horst, who encouraged and challenged me through my academic program. It was an honour and a pleasure to work with a passionate professor who sparked my curiosity about vocabulary research. Her talent and professionalism made a lasting impression on me. Thank you for your support, patience and generosity.

I would like to thank my committee members, Joanna White and Sara Kennedy, who have taken the time to offer their suggestions and corrections. Your contributions were fundamental to the success of this thesis. A thank you goes to my committee chair, Teresa Hernández González, for her support and countless words of encouragement.

Heartfelt thanks are also due to the Department of Education faculty, staff and students. I was very lucky to be able to pursue my graduate studies among very creative and energetic professionals.

I would like to specially thank the extraordinary people who have touched my life: My mother, Mariângela Martini, for being the most inspiring woman I have ever known; my father, Edson Pisani, for giving me his passion for education; and my sisters, Luciane, Cristiane, and Adriane, for motivating me with their amazing talents.

Most of all, I would like to thank my husband, Carlos Melloni, who has patiently supported and encouraged me through this process. Thanks for being such a positive and exciting presence in my life.

Table of contents

List of Tables

List of Figures

Without grammar very little can be conveyed; without vocabulary nothing can be conveyed." (Wilkins, 1972, p. 111)

Chapter 1. Introduction

My interest in vocabulary came as a natural development of my career as a second language teacher. On many occasions, textbooks available on the market were not suitable to the needs of specific groups of students I was teaching. Therefore, I had to face the challenge of developing pedagogical materials that were more relevant to their learning requirements, especially in terms of vocabulary. Those materials involved goals that ranged from basic communication to technical vocabulary for professional purposes, such as political science, environmental issues related to the natural resources industry, peace and conflict resolution, and the basic language and culture of business. In all these cases, the questions that always came to my mind were: How do I select the vocabulary for each exercise or activity? Which words are most relevant to achieving second language learners' goals?

I realized that the answer to my questions could be found in the way vocabulary is selected for the textbook itself. A corpus analysis of second language textbook materials may be useful to indicate whether experienced authors select vocabulary using word frequency lists or by simply relying on their intuitions. Such an analysis might also identify possible gaps in lexical content. Is it possible that many important words that learners need to know are missing? In addition, textbook corpora can be good indicators of the quality and quantity of language used in language classes. They provide

researchers with samples of oral and written input that serve as models for learners' production.

The research I describe in this thesis investigates the lexical content of secondary ESL textbooks used in Quebec. The reason for my choice is as follows: One of the main issues in literacy generally, and second language reading in particular, is the fact that many readers do not have a vocabulary that is large enough to understand the content of the text they read. Nation (2006) suggests that a vocabulary of 8,000 to 9,000 word families is necessary in order to understand written texts in English (p. 59). For the purposes of this study, a word family is defined as the base form of the word (e.g. *limit*) and its basic inflected and derived forms such as *limiting, limits, limitation,* and *unlimited* (Nation, 2001). The problem of such a large vocabulary size being required to read texts is more complicated in EFL contexts, where language users sometimes have very little contact with the L2 other than in school contexts. Quebec is arguably such a context. There are parts of the province where students have little or no regular contact with the English language outside of the ESL classroom. Although they have easy access to it through the media due to the proximity to English-speaking provinces and the United States, French is the language most spoken in the community (Lightbown & Spada, 1994). The distribution of languages spoken at home in Quebec is as follows: 81.1% French, 10% English, and 8.9% other language(s) (Statistics Canada, Census 2006). Thus the class textbooks may represent a large proportion of the English input many learners in this context and other similar ones are exposed to (Alsaif & Milton, 2012; Milton, 2009; Römer, 2004), and for this reason it is important to understand more about their vocabulary content.

There is reason to think that English instruction in Quebec is not adequately focused on learners' true lexical needs and that more importance should be given to vocabulary that is useful for their academic work. Cobb's (2000) investigation of English placement test data from approximately 1,000 Francophone students registered for ESL courses at a Montreal university reveals that, among other issues, Francophone learners lacked receptive knowledge of many high-frequency English words, especially those of Anglo-Saxon origin. However, research with more advanced francophone learners in Europe identified a different problem: Granger's (1996) investigation found that advanced French learners of English overuse basic Germanic words in their formal academic writing and underuse cognates. Taken together these studies suggest that there are issues related to the receptive and productive knowledge of the most frequent words in English for francophone learners. Thus there is need for research to explore high-frequency vocabulary in the pedagogical materials French-speaking secondary learners are exposed to.

The current study has two purposes: The first is to gather and analyze a corpus of textbooks to see which words Quebec secondary learners are exposed to in ESL class; the second is to use the corpus findings to develop a pedagogical word list that targets high frequency words that may need more emphasis or be missing entirely. As outlined in Chapter 3, the research methodology involved analyzing the entire lexical content of a popular series of ESL books using lexical frequency profiling software (LFP). As we will see in Chapter 4, the results of the analysis were then used to create a word list that can help materials designers and teachers in their lexical choices. But first, the next sections

explain why this project is specifically relevant to francophone students in Quebec and to students at the secondary level in particular.

Context of the Study

In contextualizing the study, the notion of text coverage is important to define. It refers to the extent to which the words in a particular text are represented on frequency lists. For example, lexical frequency profiling analysis shows that 76.32% of the words in this thesis are on the list of the 1,000 most frequent words of English according to lists based on the British National Corpus (BNC) developed by Nation (2006). About 7.73% are from the 2,000 most frequent words and 2.26% from the 3,000 for a total coverage of 86.31% by all of the 3,000. The rest are less frequent words and proper nouns.

The concept of coverage is important because one of the main issues in reading comprehension is the finding that not all learners of English have a vocabulary that is large enough to understand the content of English texts. Nation (2006) suggests that a vocabulary of 8,000 to 9,000 word-families is necessary in order to understand written texts in English at the level of 98% known word coverage; with knowledge at this level, learners will be able to understand 98% of the words in a typical text designed for native-speakers and be able to infer the meanings of the remaining 2% of unfamiliar words from context. But attaining such a large vocabulary size is probably an unrealistic goal for the population targeted in this study.

Recent testing in secondary schools in the Montreal area revealed substantial gaps in vocabulary knowledge at the 1,000 to 6,000 frequency levels (White,

Martini, Horst, & Cobb, 2012). More specifically, the mean score on a test given to learners who had completed five years of secondary schooling showed that they recognized the meanings of 59% (public school students) or 69% (private school students) of 60 items that sampled the 1,000-6,000 frequency zones. Thus about a third or more of these important high frequency words were not recognized by these learners. According to the BNC Corpus, words in the 1,000 to 6,000 frequency levels (also referred to as 1k-6k) represent approximately 88% of words in written texts, which is well below the 98% of words necessary to understand most texts (Nation, 2006). The results also suggest that students' English vocabulary knowledge improves throughout the secondary years but does not reach fully satisfactory levels to support reading comprehension of authentic unsimplified texts designed for native speakers. In other words, secondary students in Quebec are lacking a significant amount of lexical knowledge that ideally should be learned before they start their post-secondary studies.

The current study is part of a larger project called "Two paths to second language literacy: targeted word study and cross-linguistic awareness." This project is being developed at the Department of Education at Concordia University. Taken as a whole, the goal is to make sure that secondary ESL learners leave school with substantial vocabulary knowledge and are well prepared for the next stage of their education in college or university.

In brief, the goal of the study is to use corpus methodology to create a word list that is pedagogically useful to secondary level students in Quebec as a step toward addressing this gap in vocabulary knowledge. The list may be useful to teachers

who need to develop extra materials to complement the existing books, to students who will do academic studies in English, and to ESL learners who need basic proficiency/literacy in L2 (e.g. to read the newspaper in English). The following section will present an overview of corpus studies that focus on frequency lists, specialist lists, and the value of corpora for ESL.

Chapter 2. Literature Review

This chapter begins with a discussion of two key concepts that are relevant to the study: *corpus, frequency*. Then I review research that illustrates the usefulness of a corpus-informed approach to investigating the lexical characteristics of language. Lists of frequent English words are an important contribution resulting from this body of research; several of these lists are described briefly. I also present studies that show the inconsistencies between the language found in textbooks and various corpora of 'real' language. Finally, I review previous corpus-informed investigations of ESL textbooks, ending with studies specific to the vocabulary of textbooks.

Corpus Studies

A corpus is a large "collection of texts" that is compiled in order to represent a specific topic or genre (O'Keefe, McCarthy & Carter, 2007, p. 1). It can represent written or spoken discourse or both (e.g. academic discourse or technical writing for specific subject areas). Once collected, these electronic texts can be analysed in a variety of ways according to the research goal. The following paragraphs will explain the evolution and importance of corpora in vocabulary studies, some parameters to create corpora and some ways of analyzing them.

Corpus research has transformed the way scholars view vocabulary studies. Corpora started to be collected beginning in the early 1900s, first manually and later with the help of modern technological tools. The analyses of these collections of texts are now the main lexical sources for lexicography, or dictionary writing (Schmitt, 2000, pp. 68-

69). Schmitt (2010) states that corpus analysis has been one of the most important achievements in vocabulary studies which allows researchers to understand how authentic language works (p. 12). Instead of creating hypothetical models of written or spoken productions, it is possible to have access to databases containing millions of words available for a variety of quantitative and qualitative analyses. As mentioned, a corpus can contain only spoken or written texts, or a combination of both; it can also represent many variations of the same language (e.g. regional varieties of English, academic vocabulary, speech production of native and non-native speakers). Some examples of large corpora are the British National Corpus (BNC) containing over 100 million words of written and spoken British English from a variety of sources (BNC, 2010); the Cambridge English Corpus (CEC), a collection of eight corpora, containing 1 billion words from written and spoken British and American sources (CEC, 2011); the Corpus of English as a Lingua Franca in Academic Settings (ELFA), containing 1 million words of spoken ELF data (ELFA, 2007); the Corpus of Contemporary American English (COCA), a 425 million-word corpus containing written and spoken data (COCA, 2011), among others.

In order to create a corpus, it is necessary to define the criteria for text selection. These criteria should involve choosing between spoken or written productions, specific subject areas, the size of the texts, the type of discourse and the date the text was created. It is also important to make sure that the corpus is representative of the type of discourse you are trying to investigate (e.g. an academic corpus that represents different fields of university study). An example of a representative academic corpus is the Academic Corpus (Coxhead, 2000, pp. 213-214), which was carefully collected taking into

consideration various subject areas (e.g. arts, commerce, law, science) and the different types of texts used in universities (e.g. academic journals, textbooks, words from other corpora). The author also used texts written by various people in different English-speaking countries (Coxhead, 2000, pp. 219-220). The result was a corpus that has proved to be very useful to researchers interested in academic writing.

The great strength of corpora is that they show how language is actually used in different contexts. They are easy-to-access collections that allow the investigation of word frequency and lexical-grammatical profiles (e.g. collocates, chunks, syntactic patterns) (O'Keefe et al, 2007; Schmitt, 2000). For this reason, they provide researchers with the possibility of performing a variety of investigations, in order to describe specific features of authentic language. Once certain features are identified, the use of language for specific purposes becomes easier and less intuitive, hence more accurate.

In the following paragraphs, I will describe an important application – the use of corpora in creating lists of frequent word families.

Frequency

According to Schmitt (2010), vocabulary frequency is one of the most fundamental aspects of language of interest to teachers, students and textbook writers. This rests on the basic idea that the most frequent words of a language are the most useful words for learners to know. The frequency of words can be used to indicate how language behaves in specific contexts and also to compare a variety of situations. For example, frequency profiles can show the contrasts between written and spoken corpora,

native and non-native speaker corpora, and also the occurrence of specific words and multi-word units. Thus *kid* is a frequent word in spoken English (26 instance per million words in the BNC) but is much less frequent in written (7 instances per million). In a study investigating the acquisition of vocabulary through reading, Zahar, Cobb and Spada (2001) conclude that word frequency plays an important role in vocabulary acquisition (e.g. the more frequent a word is, the more likely it will be acquired). Their results show a significantly stronger influence of frequency among beginner learners, who need to encounter the same word many times in order to acquire it. The conclusion suggests that frequency is a strong predictor of L2 acquisition.

In corpus studies, vocabulary is often divided in 1,000-word frequency bands; for instance, the first 1,000 band (or 1k) represents the most frequent words in English. In this thesis, two formats are used interchangeably to refer to thousand-word frequency bands. Both figures (e.g. 1,000) and the designation 'k' (e.g. 1k) are used. Nation (2001) suggests that the 2k level represents the upper limit for high-frequency vocabulary; however, this thesis research was guided by Schmitt's (2011) definitions, which are: High-frequency vocabulary is represented by the 1k through the 3k zones; mid-frequency vocabulary includes words between the end of the 3k to the beginning of the 9k zone; and low-frequency vocabulary includes words at the 9k zone and above.

In addition to identifying frequent single words, corpus researchers have also recognized the importance of frequent multi-word units. For instance, in an analysis of the most frequent 1,000 content words in the BNC spoken section, Shin and Nation (2007) identified the most frequent collocations in English that could help L2 learners

focus on the study of multi-word units. The authors argue that collocations are specially challenging to students and that identifying the most common ones in English could be very helpful to beginner level learners. Their list of the first 100 most frequent collocations (pp. 8-10) can be very useful when teachers and material designers need to make choices of what to teach.

So far, we have considered the importance of frequent lexis learners may encounter in academic material (Coxhead, 2000), stories read in ESL class (Zahar et al. 2001), or in spoken English generally (Shin & Nation, 2007). The spoken classroom environment is another rich source for investigating the quality and quantity of vocabulary input available to L2 learners (Milton, 2009). But Brown, Waring, & Donkaewbua (2008) warn that comprehension-focused listening may only be effective for learning new vocabulary if a new word is encountered at least 30 times (as cited in Horst, 2010, p. 162), which seems to be a very high number of occurrences. Based on this parameter, Horst (2010) describes a corpus study that investigates the potential efficiency of teacher speech in the acquisition of frequent words in English. The research conducted in a Québec setting concludes that most words are not recycled as many times as needed for vocabulary acquisition; hence hearing words in teacher talk may not be a good predictor of the vocabulary students learn. The study suggests that acquisition only through listening is more complex and requires more work than just hearing a word a few times.

Nonetheless, the study points to the usefulness of a corpus-based approach. If a teacher talk corpus can help us understand a portion of input students encounter in the

classroom, it is also important to investigate the other forms of written and spoken input that L2 learners have access to (Horst, 2010; Milton, 2009). The current study will investigate the language found in secondary ESL textbooks and identify the gaps, if any, in the lexical choices made by material designers. That is, the vocabulary that occurs in the textbook will be measured against lists of words that are known to occur frequently in a large corpus that is representative of the English language as a whole in order to arrive to a useful word list for these groups of students. Thus the goal of this research is to investigate the connection between natural language corpora and textbook corpora in order to develop a principled, frequency-informed list that is specially designed to aid comprehension of written language students in Québec are likely to encounter in their studies later in high school, CEGEP, and university. Previous studies of textbook corpora will be discussed later in subsequent sections, but first we turn to research that reports on the concept of 'specialist lists'.

Specialist lists

Nation (2001) defines specialized vocabulary as lists of words based on the analysis of texts on specific and restricted topics. As examples, the author mentions special vocabulary used in technical texts, in academic articles, and in newspaper articles, among others. Specialized lexicons can include words occurring in very restrictive domains (e.g. linguistics) or broader fields of study (e.g. arts).

The importance of identifying specialized vocabulary is that the knowledge of frequently occurring specialized words can offer high levels of known-word coverage for various kinds of texts, and this makes them more comprehensible to learners (Coxhead

2000; Nation, 2001; Nation, 2006). In order to guess the meaning of possibly unknown less frequent words, learners must know the meaning of the "majority of tokens" in a discourse (Schmitt, 2000, p. 73). By way of illustration, learners whose main need is to understand chemistry texts in English may be able to reach a high level of coverage more efficiently by studying a specialist list of words that occur frequently in a corpus of chemistry texts. If the ideal coverage of a text is 98% (Nation, 2006), which is a very high percentage, knowledge of the words on special lists, added to knowledge of the most frequent words in English, can increase the accessibility of texts to ESL/EFL learners considerably.

An example of a widely used specialized vocabulary list is described in the previously mentioned work by Coxhead (2000). Her study focuses on the development of an academic word list (AWL) extracted from a corpus of 3.5 million words from 28 different academic subject areas. In order to do that, the author created lists of the words contained in this corpus and eliminated any that were part of West's (1953) General Service List because these basic words were assumed to be known to learners who are at the level of undertaking university studies in English. Other criteria used for word selection were that they had to be frequent and have range. That is, they had to occur at least 100 times in the corpus as a whole, 10 times in each of the four main-subject sections, and in at least 15 of the 28 subject areas. The problem Coxhead found in previous attempts to create frequency lists was that they were based on small corpora and not a "balanced range of topics" (p. 214). The result of her work was a new AWL containing 570 word families (e.g. *concept, conception, concepts, conceptual, conceptualisation*, etc.) that are likely to help EAP students in their studies.

Following the concept of specialized vocabulary, in the present study, I tried to arrive at a list of word families that can help Quebec secondary students develop the necessary language skills for academic literacy in English. I constructed this list by identifying the vocabulary that occurs frequently in their textbooks (i.e. a specialist textbook list) and then building on this list and supplementing it as needed. The resulting list, the SEL (Secondary ESL List) is informed by lists of the most frequent words in English and can be used by materials designers to choose lexical items to include in their materials.

In the next section, I will present a review of some studies that have explored the use of general corpora to inform textbook development. In addition, I will discuss the relevance of using textbook corpora as a tool to develop L2 pedagogical materials.

Native Speaker Corpora, Textbooks and Textbook Corpora

The importance of corpora in language learning is largely accepted in the field of applied linguistics. Römer (2004) identified a gap in research stating that most studies focus on learner corpora versus native-speaker corpora, while they should also investigate the language input presented in the classroom (e.g. textbooks, classroom speech, etc). However, the use of native speaker corpora as a tool to develop ESL/EFL materials is still a controversial subject among researchers. Although the majority of studies recognize the importance of using 'real' language in pedagogical materials (Davidson et al., 2008; Harwood, 2005; Holmes, 1988; Lee, 2006; O'Keefe et al., 2007; Römer, 2004; Shortall, 2007; Sinclair, 1997), some argue that native speaker authenticity is not a reality in the classroom context (Widdowson, 2000; Cook, 2001). In this section, I will present

the arguments based on empirical evidence that support the use of corpora, as well as the ones against it.

O'Keefe et al., (2007) and Harwood (2005) agree that most textbook language is chosen based solely on intuition about how language works in real contexts. Even though the authors recognize the purpose of using "scripted dialogues", they claim that recent evidence shows enough reasons to challenge it (O'Keefe et al, 2007, 21). Constructed and simplified texts and dialogues fail to represent the nuances and language variations that occur in 'real' writing and conversation, thus denying learners the opportunity to encounter more authentic contexts. Sinclair (1997) and Granger (1998) also believe that the use of real examples should be the basis of pedagogical materials and that they should not be created from hypothetical contexts.

In fact, Harwood (2005) suggests that the ESL/EFL materials found in the market are "highly unsatisfactory" and in need of adjustment to the most recent findings in corpus studies (p. 149). The author adds that many material developers are not familiar with language learning theories and they frequently fail to use a systematic methodology when writing textbooks. Instead of using intuition-based generalizations about language use, they should try to select suitable corpora for pedagogical purposes and use this information to produce language learning materials. Teaching/learning materials represent a substantial amount of the language learners come across (Davidson, Indefrey & Gullberg, 2008; Milton, 2009), and for this reason, they should reflect as accurately as possible what is found in large written and spoken corpora that are able to systematically capture the realities of the whole language.

The focus of studies that investigate the limitations of intuition-based textbook language can vary from single lexical items to grammatical structures, such as expressions of doubt and certainty (Holmes, 1988), the use of *if*-clauses (Römer, 2004), the subjunctive *were* vs. the indicative *was* (Lee, 2006), the use of the present perfect tense (Shortall, 2007), and as many others as research can identify. The paragraphs below will present empirical evidence of the discrepancies found between ESL/EFL textbooks and various corpora.

The first study I will discuss is Holmes' (1988) analysis of expressions of doubt and certainty, which include modal verbs (e.g. *can, must*, and *might*), lexical verbs (e.g. *appear, believe*, and *doubt*), adverbs (e.g. *actually, clearly*, and *probably*), nouns (e.g. *assumption, evidence*, and *speculation*), and adjectives (e.g. *apparent, evident*, and *likely*). The author compares the corpora of two ESL course books and two ESL reference grammars with large written and spoken corpora of natural language. The analyses indicate that the lexical content of each book varies significantly and does not contain the same proportion of occurrences we find in the corpora of natural written and spoken language. For example, although the modal verb *might* (expressing possibility) is very frequent in the actual language use (both spoken and written), it is completely disregarded in one of the books used in the study. In another textbook, there is excessive attention to modal verbs and complete disregard for the other strategies frequently used to express modality in the corpora, such as the possibilities mentioned above. The outcome suggests that some textbooks fail to teach the many possibilities to express doubt and certainty in English and therefore limit language learning to specific structures.

In a study following a similar design, Römer (2004) analysed twelve volumes of two introductory EFL course book series used in secondary schools in Germany and compared these with the spoken part of the BNC corpus. In order to obtain data representing only spoken language, the author carefully selected the textbook material, excluding narratives and other items that do not represent speech production, and compiled the German English as a Foreign Language Textbook Corpus (GEFL TC). Results show variations in the use of if-clauses between the GEFL TC and the BNC, indicating that there are discrepancies in the way we teach this structure and the way we use it. Some examples of these differences are: (1) the if-part in initial position is more frequent in the BNC; (2) the BNC presents more different combinations of tense form; (3) *if you* is the most frequent collocate in the BNC but it has the same frequency as other collocates in the GEFL TC; among others. The main issue identified in this study is that the three types of if-clauses presented in most ESL textbooks are not accurate representations of language use and that many more variations involving this structure are likely to happen in conversation. Finally, the study suggests that a review of certain structures presented in textbooks is needed and can benefit from corpus analysis.

In another study in the same vein, Lee (2006) investigated a collection of 20 grammar books and textbooks used by high school students in Hong Kong and compared their language to the Australian Corpus of English (ACE), the Australian Ozcorp, and the Hong Kong Corpus (HKC). The language feature examined is the alternation between the subjunctive *were* and the indicative *was* in the following structures: *If I was/were there* and *I wish I was/were*. The study suggests that most books analysed tend to bring a prescriptive norm to the language teaching rather than focusing on the actual use of the

varieties of English (p. 81). Results show that, although the use of subjunctive *were* is still present in writing, the frequent use of the indicative *was* is neglected in many textbooks and said to be a wrong use of the English language. These inconsistencies between textbook language and a variety of corpora indicates, one more time, the need for a complete reform in the way language is chosen for pedagogical use.

The last study presented here is Shortall's (2007) analysis of thirteen textbook series (32 books in total) and the comparison with the Bank of English 20-million word spoken British corpus (Brspok). The structure investigated was the use of the simple and the continuous forms of the present perfect. Results show that there are significant differences between the use of this verb tense in the textbook corpus and in the Brspok, such as: (1) overuse of the present perfect continuous in textbooks; (2) overuse of the adverbs *yet* and *already* in textbooks. According to this data analysis, the adverbs are said to be used as "tense markers" and give learners the wrong impression that present perfect should always be associated with them (p. 157). One more time we see how inaccurate and far from 'real' language textbook lexical choices and grammatical structures can be. Authentic language should be, at least partially, used in the design of those books investigated in the study.

Despite all the evidence showing the advantages of using appropriate corpora in textbook development to rectify problems such as those described above, some authors state that the limitations of corpus linguistics and any pedagogy based on it should be carefully delineated. Widdowson (2000) believes that corpus analysis should not define which language we teach because it describes language out of context and "only partially real" (p. 7). When we define the language content for a course through corpus studies, we

are using linguistics, which is a theoretical concept, instead of applied linguistics. We are trying to bring an isolated language feature into the classroom. The author adds that, in any case, authenticity is impossible because the classroom is a special context of its own and cannot be an authentic environment in the sense of being representative of the way native speakers use the language. Like Widdowson, Cook (2001) believes that invented examples are particularly useful to teach specific structures presented in textbooks.

After careful consideration of the empirical evidence presented in this section, I conclude that corpora are excellent references for developing pedagogical materials because they reflect the use of language in 'real' contexts. In my view, analyses of textbook corpora can bring significant pedagogical benefits to ESL/EFL learners and should be more explored in current corpus linguistics studies regardless of any arguments against their use. The range of opinions points to a possible compromise: we can try to be as authentic as possible without neglecting specific classroom needs. The proposed SEL uses a corpus approach to bridge between both realities: the textbook lexical content and what is real in the language as a whole.

In the next section, I will discuss the three key vocabulary considerations that textbook corpora can shed light on. These are the availability of high-frequency lexis in ESL textbooks, the recycling of this vocabulary in the corpus as a whole, and the recurrence of words across books (range).

Vocabulary in ESL Textbooks

On the point of the availability of high-frequency lexis, a recent study by Matsuoka & Hirsh (2010) used a textbook corpus to consider the nature of the vocabulary content of an integrated skills ESL textbook for upper intermediate learners. Results suggest that for the specific ESL textbook they investigated, there is a good chance students will learn words in the high-frequency zone (2k) but will have very little opportunity to encounter words beyond this level repeatedly. The authors suggest that supplemental materials should be provided to students in order to enhance vocabulary development in higher frequency levels and the academic level.

As for research on recycling, in a 1990 discussion, Nation shows how little repetition of words occurs in standard textbooks. For instance, in some textbooks new words are presented only once, while in others, the majority of different words are repeated only five times or less (p. 44). Similarly, Kachroo (1962) reports that Indian learners of English were able to learn most of the words that occurred seven or more times in their textbooks but fewer of the many that occurred once or twice. Also, Matsuoka & Hirsh (2010) show that 33.3% of the 2k words represented in their textbook corpus are repeated only once. Consequently, there is little opportunity for these high-frequent words to be learned.

More recently, Milton (2009) confirms that vocabulary recycling is beneficial in facilitating vocabulary knowledge. However, there is no agreement on the exact number of times each word should be repeated in a textbook. Recommendations vary from six (Zahar, Cobb & Spada 2001) or seven times (Kachroo, 1962) to 16 times

(Saragi et al., 1978) but researchers also agree that some words that are not recycled can be learned as well. In this project, I will assume that if a word family is recycled ten times or more, it is likely to be known to the learners; this is in line with the study by Horst (2009), which used this criterion in analyzing a teacher speech corpus.

On the point of range (the extent to which words recur across a number of different texts), Milton's (2009) analyses show that different textbooks designed for a particular level of proficiency vary a great deal in terms of lexical choices, vocabulary size and repetitions. The author suggests that there is little agreement on what the criteria for textbook vocabulary content should be. Although it is widely accepted that high-frequency words should be included in all textbooks, research shows that there is a large variety of vocabulary across textbooks, and thus a small number of common words in different textbooks (pp. 202-203). In a 2001 discussion, Nation explains that the distribution of words plays an important role in vocabulary acquisition. Instead of spending a long period of time studying words, these repetitions should be spread across many days. Results show that spaced repetitions and repetitions at increasingly longer intervals promote more learning (Pimsleur, 1967). As an example, if a set of words needs to be studied for thirty minutes, it is better to spread the time across many days in larger intervals that to study these words for thirty minutes in only one day. Although there is no agreement on the exact intervals for studying a word, research shows that spaced repetitions enable learners to remember words for longer periods of time.

Taken together, the studies presented here reinforce the idea that the most useful words for teaching a language are the ones with high frequency and wide range across the

materials (Richards, 2001). The proposed research aimed to follow the above recommendations; it provides secondary ESL teachers in Quebec with a word list that may be used as an additional source of vocabulary to complement existing Ministry-approved textbooks.

Summary of the Literature Review

In the discussion of research above, I first established the usefulness of corpora in capturing the nature of authentic language, and I described an important corpus application: the identification of frequent words that are important for language learners to know. Secondly, I showed that corpus-based frequency lists geared towards particular learner needs can help make learning efficient. In the case of secondary ESL learners in Quebec, if we can focus our teaching on words that cover the majority of written and spoken productions, we would improve syllabus design significantly. However, it is probably not possible to ensure that these learners can acquire the 8,000 or 9,000 high frequency words needed. Schmitt (2011) offers a useful way forward. He suggests that, in terms of cost/benefit, learners should be taught high-frequency vocabulary (1k through 3k) explicitly. Therefore the proposed research focused on the occurrence of words in these three frequency bands in a series of ESL textbooks.

Previous corpus-informed studies of language textbooks examined in the literature review showed that there are gaps between real English and the English of ESL textbooks. Studies of particular features (such as if-clauses or the perfect tense) found there was over- or under-emphasis of the feature. Especially relevant for my research are the studies that considered all of the vocabulary presented in a particular textbook (or set

of textbooks) in terms of the inclusion of high frequency words, the number of times these words were recycled, and their spread or range across the chapters or units of the textbooks. The picture that emerges consistently shows deficits on all three points. Given these findings, it is likely that the textbooks I analyse will also show that some word families from the 1k, 2k and 3k lists are missing entirely and those that do occur are not recycled at least ten times over three of three books. But none of the research I encountered set out to address such deficits positively by providing a principled, research-informed list specially designed to meet the needs of a particular group of learners. This is the goal of the thesis research which evaluates the vocabulary strengths and weaknesses of an entire series of ESL books, resulting in a pedagogical list that will supplement any deficits that are found.

Therefore, after investigating the current research and identifying the need for studies involving textbook corpora and the use of 'real' language in pedagogical materials, I propose a new corpus study and a list of word families in English, the SEL, which was designed to help Quebec secondary students in their academic work. My intention is not to fault the work of material designers when developing a textbook series, but to find useful solutions to make their job easier in the future. In the next section, I will present the research questions proposed for this study.

Research Questions

The main goal of this study is to arrive at a useful pedagogical list, the SEL, which may help secondary students in Quebec in their future academic studies. The objective is to form a list containing between 500-1,000 word families. This size was

chosen because it represents a manageable number of words to target in a program of ESL instruction in the three years of high school that follow on Secondary 2 (i.e. about 10 new words per week x 35 weeks x 3 years). Although Nation (2001) suggests the 2k level as the upper limit for high-frequency vocabulary, this study followed Schmitt's (2011) more recent parameters for high-frequency vocabulary: the 1k-3k zones. Due to the importance of teaching high-frequency lexis explicitly (Nation, 2001; Schmitt, 2011), this pedagogical list will contain only words included in the 1k, 2k, and 3k levels of the BNC Corpus. However it will not include 1k, 2k, and 3k words that have been recycled repeatedly in the materials since these words can be assumed to be already known. Additionally, an analysis of the occurrence of words between the 3k and the 9k frequency zones (mid-frequency vocabulary) will be included because it will provide teachers and materials designers with an idea of the next vocabulary levels they should focus on. Details of the selection process follow below and in the methodology section.

I propose the following research questions:

Research question 1:

a. To what extent are all of the 1,000, 2,000 and 3,000 most frequent word families of English represented in a widely used sequence of secondary textbooks?

b. Does this change with the inclusion of the workbooks?

Research question 2:

a. To what extent are these words used repeatedly across the books (in all three books) and in the entire corpus (10 times or more)?

b. Does this change with the inclusion of the workbooks?

Research question 3: Which 1,000, 2,000, 3,000-level words are missing in the core textbooks?

Research question 4: To what extent are all of the 4,000, 5,000, 6,000, 7,000 and 8,000 most frequent word families of English represented in the core textbooks?

My hypotheses for each question are:

Research question 1:

a. High frequency words are underrepresented in the vocabulary content of the book series.

b. The inclusion of the workbook content does not change the frequency distribution of words.

Research question 2:

a. Most words are not recycled more than 10 times and do not occur in all three books.

b. The inclusion of the workbook content improves the recycling of the words.

Research question 3: Many 1,000 to 3,000-level words are missing in the books.

Research question 4: Most 4,000, 5,000, 6,000, 7,000 and 8,000-level words are missing in the books.

The purpose of the first question that investigates the 1k, 2k and 3k lexis that occurs in the secondary textbooks is to determine the extent to which learners are able to meet these important words. It might be expected that by the end of three years of secondary-cycle 2 ESL study, learners would have met all 3,000 of these frequent word families in their learning materials at least once. However, the textbook analysis by Matsuoka and Hirsch (2010) showed that only 603 families of the 1,000-member 2k list occurred in the materials they investigated. Similarly, a classroom speech corpus by Horst (2010) showed that only 583 families of the 2k level and 309 families of the 3k level occurred in the 121,000-word teacher talk corpus. In the same vein, an analysis of a typical series of Quebec primary ESL books found that high-frequency words are underrepresented (about 26% of the 1k, 2k, and 3k levels were missing) in the corpus (Horst, White, & Cobb, 2011). These findings are the basis for the expectation that many high frequency words will be present in the materials but the 1k, 2k, and 3k lists will not be represented in their entirety. This previous research is the basis for the hypothesized outcome to the first research question. The workbooks are supposed to contain the same words found in the textbooks; so adding the workbook material to the textbook corpus is not expected to substantially change the extent to which high frequency words occur.

As for the second question that pertains to repetition, while it cannot be assumed that learners know all of the words that occur in a textbook, it is reasonable to think that words that recur frequently are likely to be known (Kachroo, 1962; Saragi et al., 1978; Zahar et al. 2001). It is expected that there will be a core set of high frequency words that are recycled often in the textbooks and these will be considered as 'known'. It is important to identify 'known' words so that they can be excluded from the proposed pedagogical list. However, a study by Horst (2010) shows that most words ranging from the 3k to the 20k levels may occur only once or twice in an entire 121,000-word teacher talk corpus representing more than 30 hours of class time. Similarly, in a previously mentioned study, 33.3% of words from the 2k frequency level occurred only once in a textbook corpus representing an entire book (Matsuoka & Hirsch, 2010). So there is reason to think that many words in the targeted 1k-3k zones will not be repeated ten times or more in the textbook series investigated here; this is the basis for Hypothesis 2a above. The addition of the workbooks, which are supposed to be used to complement and reinforce the textbook content, is expected to improve the recycling of the words in the corpus.

The third question concerns 1k, 2k, and 3k word families (using Nation's 2006 BNC-based lists) that are missing entirely. These are important to include on the proposed SEL. After analysing the studies presented above, there is reason to think that there will be many missing words in the high-frequency zones in the textbook corpus. These words will be identified and selected to be included in the new word list.

The fourth research question deals with the word families from the 4,000 through the 8,000-level. According to Nation (2006), these are words that (along with the more basic 1,000, 2,000 and 3,000-level words) provide a good level of coverage of unsimplified texts in English. Therefore, they are essential to help students develop their reading comprehension. It is reasonable to think that some words from the 4,000 to 8,000 levels will occur, but due to the limits on what can reasonably be included in three textbooks, the corpus is not expected to contain all the words at these levels (as stated in the hypothesized outcome for question 4 above). Those that do occur are not expected to be recycled often enough to be learned by secondary ESL students.

Ideally, Quebec learners would complete their secondary ESL courses with knowledge of more than 3,000 high frequency words. All of the 8,000 or 9,000 frequent English words that Nation (2006) has identified as being important for effective reading of unsimplified English texts would be an even better aim. However, as mentioned previously in this thesis, this seems to be an unrealistically high goal. In a study that involved testing over 1,000 students on a measure of receptive vocabulary size, White *et al.* (2012) found that at the end of their secondary ESL training, there were still substantial gaps in learners' knowledge of words at the crucial 1,000, 2,000 and 3,000 levels. Therefore, it was decided that the pedagogical list should definitely include any of the 1,000, 2,000 and 3,000 most frequent words of English that do not recur frequently in the learning materials. Recently, Schmitt (2011) has also emphasized the importance of mid-level vocabulary between the 3,000 and 9,000 most frequent zones. If the textbook corpus analysis identifies only a handful of missing words at the 1,000, 2,000 and 3,000 levels, then the bar should be raised to include the words from the 4,000 to the 8,000-

levels. As mentioned, the goal was to create a list of several hundred 'most needed' words that can eventually serve as the focus of supplementary vocabulary learning activities.

Chapter 3. Methodology

In this section, I will present the methodology that was adopted to carry out this research project. It is divided into two parts. First, I will describe the materials used in the analysis. Second, I will outline the procedures for gathering the corpus and for making the word list.

Materials

Textbook Selection

For the purposes of this study, I have chosen a series of secondary ESL books used in Quebec. The series is included in the 2011/2012-list of pedagogical materials approved by the Ministère de L'Éducation, du Loisir et du Sport Québec (MELS) to be used in ESL classes in schools in the province. In addition, these books were identified as being used by three of the seven schools visited in a recent data collection in the Montreal area (Horst et al., 2011).

The series chosen is *Collection Quest* for cycle 2 published by Chenelière Éducation. It contains three volumes (for secondary levels 3, 4 and 5) with a variety of activities including reading, grammar, listening and tasks which were specially designed for secondary students in Québec. Each of the three textbooks has an accompanying workbook. In addition to the texts of the learning activities presented in the books, the corpus includes headings, titles, instructions for tasks, notes, examples, dialogues, fill-in-the-blank exercises, and the written scripts of listening activities that are also provided in written format. For practical reasons, other audio materials (e.g. DVDs and CDs) and the

teacher's guide were not included in the materials selected for the corpus; in other words, the corpus reflects written textbook and workbook material. It is recognized that this provides a comprehensive but not entirely complete picture of the vocabulary input that users of the textbook series are exposed to. For the purpose of this research, both the textbooks and the workbooks were analysed. However, only the textbooks were used to select the final word-list. The reason for this decision was the finding that the majority of secondary ESL teachers I surveyed for a study by Horst *et al*. (2011) use only the textbooks in their classes.

Corpus design

The SEC (Secondary ESL Corpus) contains all of the running words in the three volumes of the secondary-level textbooks and workbooks. The corpus is divided into sub-corpora at three levels: the book chapter, the book, and the textbook plus the workbook (Figure 1). The sub-corpora are organized as follows:

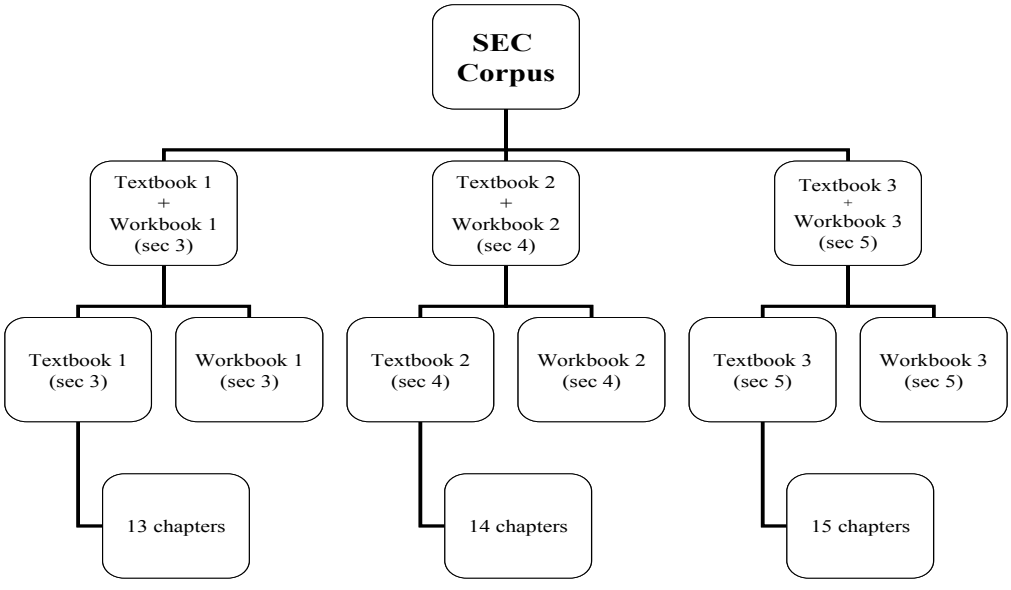

Figure 1: Corpus composition with sub-corpora levels.

Procedures

Corpus collection

In order to gather the SEC identified above, the content of the textbooks was scanned using the software ABBYY Fine Reader 5 Sprint Plus, which allows images to be saved in .txt or .doc extensions. Due to possible distortions related to the images in the textbooks, some sections or chapters were typed using Microsoft Office Word Processor software.

Each chapter was saved in .doc extension and edited for spelling and other possible distortions caused by the scanning process and/or typing. Subsequently, each file was saved in text format (.txt extension), completing the first level of sub-corpora in the study (the unit/chapter level).

To facilitate the creation of the sub-corpora and the corpus at the next levels (e.g. the book level, the textbook plus workbook level, and the SEC level), each text file was entered into the text-tool *Corpus Builder*, available at Lextutor. This website (available at www.lextutor.ca) offers a variety of software tools for processing corpora. The software joined up the components of each file, built each of the subsequent corpora, and finally joined these together to form the larger textbook corpus (SEC).

Analyses and list building

The frequency analysis in this study was done using lists based on the British National Corpus (BNC). The reasons for choosing a corpus of British English and not a corpus of North American English for this research are as follows: First, the tools freely

available for corpus analysis offer frequency profiling software that uses the BNC as its base; second, there is great overlap of high-frequency words in frequency lists based on different corpora (Nation, 2001), which make them very stable. In addition, the BNC lists are based on an updated corpus and allow for more nuanced frequency distinctions than previous programs (e.g. Nation & Laufer's earlier 1995 version of *Vocabprofile*).

In order to answer research questions 1 and 4 about the extent to which the 1,000-8,000 most frequent words of English are found in the textbooks, the SEC and each of the sub corpora was entered in the *Range* software (Heatley, Nation, & Coxhead, 2002). The software identified the frequency of each word in 1,000-word bands based on the British National Corpus (Nation, 2006). The *Range* software shows how the most frequent words are represented in the textbooks by providing the frequency profile of the corpus content. That is, the *Range* output shows the number of textbook families that are on the list of the 1,000 most frequent English word families according to the BNC, the number on the list of the 1,001-2,000 most frequent, the 2,001-3,000 and so on. The output also provides extra information about mid-frequency, low-frequency and off-list words. As an example of the output, Table 1 below shows the vocabulary profile of a section of this thesis with numbers of families at each frequency level shown in the last column.

Table 1

Frequency Distribution for a Section of This Thesis

Freq. level	Tokens/%	Types/%	Families
1k	993/78.13	301/65.86	229
2k	112/ 8.81	76/16.63	60
3k	31/ 2.44	13/ 2.84	11
4k	38/ 2.99	16/ 3.50	15
5k	12/ 0.94	6/ 1.31	5
6k	1/ 0.08	1/ 0.22	1
7k	7/ 0.55	5/ 1.09	5
8k	8/ 0.63	3/ 0.66	2
9k	1/ 0.08	1/ 0.22	1
10k	4/ 0.31	2/ 0.44	1
11k	0/ 0.00	0/ 0.00	0
12 k	1/ 0.08	1/ 0.22	1
13k	2/ 0.16	1/ 0.22	1
14k	9/ 0.71	2/ 0.44	2
15k	15/ 1.18	6/ 1.31	6
not in the lists	37/ 2.91	23/ 5.03	?????
Total	1271	457	340

As for research question 2, which pertains to the recycling of words across the books, the analysis involved entering each sub-corpus into the *Range* software. The goal was to determine whether the 1k, 2k and 3k words that are found in the books recur often enough for them to be considered as known. We have a series for three years of school, with a total of three books. The first criterion was to verify if the words occur in all three books, providing students with spaced repetitions of these words. By way of illustration, Table 2 shows words that occur in five individual pages of this thesis. Note that *'vocabulary'* and *'lexical'* occur in each of these pages (designated in the output as F1, F2, F3, F4, and F5), while *'profile'* occurs in only two of the pages and *'because'* in only one of the pages.

Table 2

Range – Partial Results

Freq. level	Families	Range	Freq. Family	F1	F2	F3	F4	F5
1	because	1	1	0	1	0	0	0
1	english	4	14	1	4	4	5	0
2	knowledge	3	7	3	2	2	0	0
4	vocabulary	5	18	2	4	2	4	6
4	profile	2	2	0	1	1	0	0
5	corpus	4	8	3	1	2	0	2
8	textbook	3	7	0	0	1	2	4
10	pedagogy	3	4	1	0	2	0	1
14	lexical	5	8	1	1	3	2	1

Subsequently, I had to verify whether these 1k, 2k and 3k words were recycled at least ten times (or more) in the entire corpus. The same *Range* software was used for this analysis. In addition to the totals shown in Table 1, *Range* output also lists all word families in a submitted text (in this case, the entire corpus). These are grouped by frequency list (1k, 2k, 3k, etc.) with numbers of occurrences in brackets. To illustrate, here is a sample of 3k families from a portion of this thesis.

BNC-3,000 families: [fams 42 : types 54 : tokens 167]

approximate_[2] artificial_[1] author_[14] behave_[1] canada_[1] capture_[1] cart_[1] chunk_[1] compile_[2] comprehensive_[1] contrast_[1] convey_[2] core_[1] electronic_[1] expose_[3] extract_[1] frequent_[81] grammar_[3] illustrate_[1] invent_[1] manual_[1] multi_[3] neglect_[2] norm_[1] oral_[1] origin_[2] outcome_[1] portion_[1] predict_[2] quantity_[3] reveal_[2] scope_[1] sole_[1] speculate_[1] supplement_[3] tense_[5] thus_[5] token_[1] transform_[1] upper_[2] versus_[2] zone_[6]

The result of this analysis is a list of 1k, 2k, 3k words that are assumed to be known by the textbook users. These were excluded from the SEL.

Answering the third research question about the high-frequency words that are missing in the textbooks involved developing a word list based on the previous results. In order to build the word list, I included any missing 1k, 2k, and 3k families on the BNC list; i.e. families that do not occur in the textbook corpus at all. I also included any 1k, 2k, and 3k families that occur in the materials but are not considered known according to the

criteria used to answer the second research question. That is, any words from these frequency levels that occur in only two textbooks (or fewer) or nine times (or fewer) in the corpus as a whole were considered as words in need of further instructional focus, and were therefore included on the SEL.

To identify the missing words, I used the *Vocabprofile* program at Lextutor. This program allows the user to identify the frequency profile by families and also the words that are missing in each frequency level. In my analysis, the 'text' consisted of all of the words on the 1k, 2k and 3k BNC frequency lists found in the textbook corpus. The option *VP-Negative* was selected and the software provided me with a list containing all the 1k, 2k and 3k level words that are missing entirely in the books. Of interest are the BNC words that are unique, i.e. the items that do not overlap with the known words. These words plus the under-represented words at the same frequency levels make up the SEL.

The procedure was expected to result in a list of only several hundred words, but as we will see, the number of 'missing' words proved to be very high. This means it was not practical to also include the 4,000-8,000-level words in this secondary pedagogical word list. If the inclusion of these mid-frequency words had proved feasible, the list could have done an even better job of meeting the goals for secondary students aiming to pursue CEGEP and University studies. However, we have opted in favour of practicality.

Finally, the software Familizer, available at Lextutor, was used to produce a full family list for all the words in the list. Thus the family for headword *'symbol'* will include *symbolic, symbolism* and *symbolize.*

Chapter 4. Results

In the following chapter, the four research questions will be answered. The first research question addresses how well the most frequent words in English (1k-3k) are represented in the books. The second research question deals with the recycling of these words across the books and in the whole corpus. The third research question identifies high-frequency missing words in the books. Finally, the forth research question focuses on the occurrence of the words from the 4,000 to the 8,000 frequency levels, which offer L2 reader, according to Nation 2006, the minimum vocabulary level needed for reading comprehension of unsimplified texts in English (in addition to the 3,000 frequent words needed for basic comprehension).

RQ 1a: To what extent are all of the 1,000, 2,000 and 3,000 most frequent word families of English represented in a widely used sequence of secondary textbooks?

The textbook corpus is composed of 283,299 running words and a total of 7,037 word families. In terms of high-frequency words, the frequency distribution (Table 3) shows that over 98% of the 1k word families occur in the books. A similar result is found at the 2k level, with the occurrence of 95% of families in the corpus. At the 3k level, the number of families represented in the books is over 85%. The results show that most of the high-frequency word families are represented in the books, with a minor decline in occurrence at the 3k level. Therefore, my first hypothesis that high frequency words are underrepresented in the vocabulary content of the book series is not confirmed here. There are some deficits, but they are not large.

Table 3

Frequency Distribution of the Words in the Textbooks – 1k to 3K

Freq. level	Tokens/%	Types/%	Families
1k	224535/79.26	3486/22.11	989
2k	22997/ 8.12	2591/16.44	950
3k	8067/ 2.85	1693/10.74	856

RQ 1.b.: Does this change with the inclusion of the workbooks?

The textbook corpus combined with the workbook sub-corpora resulted in a total of 364,342 running words and about 7,566 word families. The frequency distribution of the 1k, 2k and 3k level words (Table 4) does not show substantial differences after the addition of the workbooks to the corpus. At the 1k level, only one additional word is found in the workbooks; at the 2k and 3k levels, 9 and 20 new words occur in the workbooks, respectively. Since the workbooks are expected to recycle and reinforce the content of the textbooks, the results shown in Table 4 are not surprising. The results confirm my hypothesis for this research question which is that generally, the use of workbooks does not greatly increase the numbers of 1k, 2k and 3k words the textbook users are exposed to.

Table 4

Frequency Distribution of the Words in the Textbooks and Workbooks – K1 to K3

Freq. level	Families (Textbooks)	Families (Textbooks & workbooks)
1k	989	990
2k	950	959
3k	856	876

RQ 2.a.: To what extent are these words used repeatedly across the books (in all three books) and in the entire corpus (10 times or more)?

The recycling of high-frequency words in the SEC changes considerably from the 1k to the 3k levels (Table 5). Most of the 1k level words (94%) occur 10 times or more in the corpus, while only 65% of the 2k words are repeated 10 times or more. At the 3k level the recycling of words is even lower; only 27% of these words are repeated 10 times or more. There is a substantial decline in the recycling of these words at the 3k level; consequently, opportunities to learn these words are not well supported.

For the purposes of this research, a word is considered well represented in the corpus if: (1) it is repeated 10 times or more in the corpus; and (2) it occurs in all three books (spaced repetitions). Table 6 shows the percentage of words in the core textbooks that reach these criteria. Out of 989 1k words found in the books, 908 meet the criteria.

There is a large difference in the recycling of the 2k and 3k words in the corpus. Out of the 950 2k words in the corpus, only 560 meet the criteria. Out of 856 3k words in the corpus, only 190 are recycled 10 times or more and occur in the three books. The results partially confirm my hypothesis that most words are not recycled more than 10 times and do not occur in all 3 books.

Table 5

Words Recycled in the SEC (Secondary ESL Corpus)

Freq. level	Families	Less than 10x	10x or more
1k	989	57 (6%)	932 (94%)
2k	950	328 (35%)	622 (65%)
3k	856	624 (73%)	232 (27%)

Table 6

The Recycling of Words (repeated 10x or more) Across the Books

Freq. level	Words repeated 10x or more	Repeated in 3 books	Repeated in 2 books	Repeated in 1 book
1k	932	908 (97%)	24 (3%)	0%
2k	622	560 (90%)	60 (9.7%)	2 (0.3%)
3k	232	190 (82%)	39 (17%)	2 (1%)

RQ 2.b.: Does this change with the inclusion of the workbooks?

The inclusion of the workbook content to the corpus brings a minor improvement in the recycling of the words. In terms of the number of times a word is considered well represented in this research, the 1k level shows an increase of 26 words with the addition of the workbooks. At the 2k and 3k levels, the addition of the workbooks brings 69 and 61 extra words, respectively, that are repeated 10 time or more in the entire corpus (table 7). The results show that only 156 more words out of 3,000 (1k, 2k and 3k levels) reach the first criterion defined in this research. It does not fully confirm my hypothesis that the addition of the workbook content improves the recycling of the words.

Table 7

Words Recycled 10 times or more in the SEC (textbooks plus workbooks)

Freq. level	Families (textbooks)	10x or more (textbooks)	Families (textbooks + workbooks)	10x or more (textbooks + workbooks)
1k	989	**932**	990	**958**
2k	950	**622**	959	**691**
3k	856	**232**	876	**293**

RQ 3. Which 1,000, 2,000, 3,000-level words are missing in the core textbooks?

Given that most secondary ESL teachers do not use the workbooks in class, only the content of the core textbooks was used in this analysis. The total number of missing words by level is as follows: 1k level - 18 words; 2k level – 65 words; and 3k level – 158 words. These numbers represent the total number of missing words without the exclusion of proper nouns and words mostly used in British English. The exclusion of these words was done later during the selection of the SEL (Secondary ESL List). The results partially confirm my hypothesis that many 1,000 to 3,000-level words are missing in the books.

RQ 4. To what extent are all of the 4,000, 5,000, 6,000, 7,000 and 8,000 most frequent word families of English represented in the core textbooks?

Table 8 shows the frequency distribution of the words from the 4k level to the 8k level. The occurrence of these words in the textbooks decreases substantially if compared to the 1k, 2k and 3k levels. At the 4k and 5k levels, more than half of the words occur in the books, but there are still many words missing. At the 6k level, less than half of the words occur in the books. At the 8k level, only about 29% of these word families are present in the textbooks. The results show students have very few opportunities to encounter these words in the materials. My hypothesis that most 4,000, 5,000, 6,000, 7,000 and 8,000-level words are missing in the books is confirmed.

In terms of the recycling of these words, few of them reach the first criterion used in this research (words repeated 10 times or more). Table 9 shows that most of the words at these frequency levels are repeated fewer than 10 times in the entire corpus, leaving

few opportunities for students to encounter them in the materials. Out of the 5,000 words from the 4k to the 8k levels, only 245 (5%) are recycled 10 times or more in the corpus. Tables 8 and 9 show that the vocabulary needed for reading comprehension of unsimplified materials is not well represented in the textbooks.

Table 8

Frequency Distribution of the Words in the Textbooks – K4 to K8

Word list	Tokens/%	Types/%	Families
4k	5656/ 2.00	1207/ 7.66	703
5k	2952/ 1.04	828/ 5.25	564
6k	2183/ 0.77	619/ 3.93	460
7k	2140/ 0.76	438/ 2.78	334
8k	801/ 0.28	356/ 2.26	292

Table 9

Words Recycled in the SEC (Secondary ESL Corpus)

Freq. level	Families	Less than 10x	10x or more
4k	703	573 (82%)	129 (18%)
5k	564	516 (91%)	48 (9%)
6k	460	428 (93%)	32 (7%)
7k	334	309 (93%)	25 (7%)
8k	292	281 (96%)	11 (4%)

To summarize, the results presented above show that there is a deficit in the vocabulary content of the textbooks. The 1k level is well represented in the books, but starting at the 2k level, there are deficits in the occurrence and recycling of these words. Most words at the 2k level are present in the books; however, many of them are not adequately recycled. There is also a clear a deficit in the representation of the 3k level words in the book, particularly in the recycling of most of these words. In addition, there is a substantial gap in the representation of the important words from the 4k to the 8k levels.

Among the high-frequency levels (1k-3k) discussed here, the 3k is the one that requires more attention. In order to fill in the gap in lexical content, a list of 1,281 word families was developed following the methodology previously described. This list can be used to guide the development of supplemental materials that will give support to the learning of high-frequency words in English. For a complete supplemental list of 1,281 word families, see appendix A.

Chapter 5. Discussion

In this chapter, I will discuss the results of the analysis of the secondary Quebec ESL Quebec Textbook Corpus (SEC). The two versions of this corpus contain 283,299 running words (textbooks only) and 364,342 running words (textbooks plus workbooks). It was compiled from three ESL textbooks used in the three last years of the secondary level in many schools in Québec. I will explain the deficits found in the current materials in terms of high-frequency words, mid-frequency words, and their implications for the learning of these words. I will also give more details about the pedagogical word-list that can be used as a supplement to the current books.

The High-Frequency Levels

Researchers widely agree that high-frequency vocabulary must be taught in ESL/EFL classrooms, and that learners benefit from learning it (Schmitt, 2011; Nation, 2001). Nation (2001) emphasizes that anything we can do in order to teach students the most frequent words in English is highly beneficial to their lexical knowledge. Based on the principle that vocabulary frequency is one of the most fundamental aspects of language of interest to teachers, students and textbook writers (Schmitt, 2010), we could assume that ESL textbooks focus a great deal of their content on these words. Although the results of the study show that most of the words in the textbooks belong to the high-frequency word bands, there are important issues with the way they are represented in the books.

At the 1,000-level, 989 words occur in the textbooks (Table 3) and 908 of these are recycled 10 time or more and in the three books (Tables 5 & 6). The inclusion of the workbooks does not change this picture, showing that its use does not increase the number of 1k words textbook users are exposed to. But what kinds of words form the 1k-level included in the SEC? Many of them are function words, the following being the most frequent ones in the corpus: *'the'*, *'to', 'and',* and *'a'*. They occur in the corpus respectively 15,113 times; 8,466 times; 6,892 times; and 7,745 times. There are also many content words such as '*people', 'work', 'life', 'day',* and *'music'*, among others. Although they are very basic English words, it is important to mention that secondary students in Québec need to increase their vocabulary level in all high-frequency bands, even the very basic 1k. Work by White *et al* (2012) showed that the mean number of 1k words known by Quebec students in their last year of secondary was incomplete, amounting to about 800 families. The investment in learning all of the 1k families is worth making. The knowledge of these words will provide students with approximately 70-75% known text coverage (Nation, 2001), which is still far from the ideal 98% known text coverage, but a substantial step towards it. It is clear that the books investigated here represent the essential 1k-level very well, providing students with the necessary exposure to these words. But the finding that the secondary students investigated by Horst *et al* (2011) did not know the full 1k list suggests that exposure alone may not be enough to ensure that this important vocabulary is learned. I return to this point in the section on pedagogical implications.

As for the next frequency levels, the 2k and the 3k, the representation of these words in the three ESL books changes considerably. Most words at these levels occur in

the books (950 at the 2k level and 856 at the 3k-level), showing that very few words are missing entirely in the corpus. Some examples of the most frequent 2k word families in the SEC with the number of occurrences in brackets are: '*text'* (695), '*below'* (207), and '*animal'* (210). Among the most frequent 3k word families in the SEC with the number of occurrences in brackets we find: *'grammar'* (180), *'adventure'* (128), and *'invent'* (174). The inclusion of the workbooks does not change this picture in a way that was expected by my hypothesis. Only 156 extra words were considered well represented in the books, which is a very low improvement in the number of words that occur and in their repetition across the books. In addition, we learned that most ESL teachers that use these books reported that they do not use the workbooks as part of their pedagogical materials.

Although most of the 2k and 3k level words occur in the books, the number of repetitions and their recycling across the books do not meet the requirements to be considered well represented in the books. Only 560 2k words and 190 3k words are recycled 10 times or more and in the three textbooks. As examples, a 3k word such as *'intervene',* occurs only once in the corpus; a 3k word such as *'retain'* occurs only twice in the corpus and in only one of the three books. It means that students will be able to encounter these words in their books only once or twice during the three last years of secondary school. Richards (2001) stresses that words in language samples are relevant to language learners only when they are presented frequently in a wide range. For this reason, I conclude that most of the 2k and 3k words need to be made more available to students using these textbooks.

There is clearly a deficit in the occurrence of the high-frequency words in the books. If we count all the missing and the underrepresented words, there are approximately 1,281 words out of 3,000 that students will not encounter or will rarely encounter in their textbooks. What is the implication of this deficit for their ability to understand written texts in English? Coverage figures for the 3k vary somewhat depending on frequency lists and corpora used by different researchers. Nation's (2006) work using more recent BNC-based lists found between 89% and 95% coverage for various text genres. For this reason, if students have difficulties understanding the meaning of these basic words, they will have problems with basic reading comprehension of texts in English. Although knowledge of the 1k, 2k and 3k word families in English does not meet the 98% coverage proposed by (Nation 2006), they clearly form an important list of words to know; it is also a realistic goal for the three last years of secondary studies.

The Mid-Frequency Level

As previously mentioned, the ideal known word coverage necessary to understand and enjoy reading text in English is approximately 98%, which means a reader should have a vocabulary as large as 8,000 to 9,000 most frequent words in English (Nation, 2006). The analysis of the SEC showed that words beyond the 3k level are not well represented in the textbooks or do not occur at all. A closer analysis of words from the 4k to the 8k levels showed considerable deficits in their occurrences. As tables 8 and 9 show, even when the words occur, they are not recycled consistently in the books. Out of the 5,000 words included in these frequency levels, only 5% (245) are recycled more than

10 times. Examples of mid-frequency words that occur only once in the entire corpus are: *'precaution'* (4k); *'drawback' (5k)*; *'deception'* (6k). It is unreasonable to think that all the words from the 1k to the 8k levels would be repeated ten times or more and in all three books, but there is clear evidence that the mid-frequency level words were not a priority when these books were designed. These results are similar to previously mentioned textbook corpus research that showed very limited learning possibilities beyond the 2k level (Alsaif & Milton, 2012; Matsuoka & Hirsh, 2010).

The great deficits in the mid-level vocabulary are not exclusive to the textbooks used in secondary schools in Québec. It is not surprising that the words at the 3k frequency level and the lower frequency levels are not recycled as often as the 1k and 2k vocabulary. In an investigation of a variety of written texts commonly encountered by ESL learners (press, fiction and academic), Cobb (2007) found that few words beyond the 2k level are recycled well enough to be learned. The author concludes that, in order to deal with these important words, it is necessary to use supplemental materials that could increase learners' exposure to them. This is precisely the kind of need the SEL is intended to address.

The issue of teaching high-frequency words is well established and all of the textbooks investigated in this study show that there are attempts to represent these words in pedagogical materials. However, the mid-level vocabulary represents a frequency zone that has received very little attention from materials designers (Schmitt, 2011). How can we teach all these words within the time constraints of a language class and the limits of a textbook? My suggestion here is that first, we should focus on the best ways to effectively teach high-frequency words. Second, we need to start thinking about the mid-

frequency vocabulary as an essential tool for reading comprehension. If time and materials limit the teaching of all these words, a selection of the most relevant words to specific learners should be made (e.g. words from the 4k to the 8k levels that are used in the academic setting). These are ideas that I will return to in the section on pedagogical implications.

The SEL

After analysing the SEC and identifying the gaps in the occurrence and recycling of the high-frequency English words, this study proposes a word list that contains the missing and underrepresented words in these bands. All proper nouns were also removed. The reasoning for this is as follows: Some of the proper nouns are parts of lexical sets that are likely to be already known (e.g. *Thursday* and *April*) to secondary learners of English through previous study. Others such as the names of countries (e.g. *Canada* and *New Zealand*) can be considered to be lexically transparent and not really items that need to be taught. In addition, words that were identified as mostly used in British English such as *'bloke'* and '*quid'* were eliminated from the final SEL. The selection of 'British words' was made with the help of a native speaker of English (North American variety). The idea is neither to replace the existing textbooks nor to challenge the quality of their content. The SEL can be used to guide teachers in their lexical choices for extra activities. If teachers are able to make all these words available to learners in the classroom, the quality of the classroom input can be improved considerably.

It is important to remember that the context of our research is the secondary level in Québec, which is arguably an EFL context (Lightbown & Spada, 1994). When

learners' main contact with the English language is in the classroom environment, it is important to make sure that this input is large enough and of good quality. Two important studies show the need for a more extensive attention to these words in Québec schools: first, the substantial gaps in vocabulary knowledge at the high-frequency levels identified among secondary students (White et al, 2012); second, the lack of receptive knowledge of high frequency words indentified in university entrance level tests of Francophone students (Cobb, 2000).

A new ESL word list was created to address the deficits in the high-frequency vocabulary in the current pedagogical materials. The goal is to assist teachers in their word choices when they choose, adapt, or develop new activities to be used in the classroom. All 1,281 headwords of the SEL can be found in Appendix A. A sample of the list with complete families for each headword is shown in Appendix B.

Chapter 6. Implications and Conclusions

To summarize, the main findings of this study show that there are considerable deficits in the vocabulary presented in the three books of the *Collection Quest.* These gaps are related to the frequency of words occurring in the books and their range across the materials. The 1k level was considered well represented because most words at this level occur in the books frequently and are widely recycled across the volumes. At the 2k and 3k levels, most words also occur in the books; however their frequency and range of occurrence is not satisfactory to promote vocabulary learning.

In order to address the issue of making these very important words available to learners, the current study aimed to create a new English word list to complement the lexical content of ESL textbooks used in secondary classrooms in Quebec. The idea was to give these students more opportunities to encounter the 3,000 most frequent words in English, providing teachers with a list of selected words that should be taught explicitly. The result was a list containing 1,281 high-frequency word families (appendix A).

As for the mid-frequency vocabulary, results show that students have very few opportunities to encounter these words in their books. Most of the words between the 3k and the 9k levels are not recycled frequently in the corpus. Although there is a clear deficit in vocabulary at these frequency bands, it was not practical to include them in the SEL. There were already too many important high-frequency words that should be made available to learners, so the inclusion of more words in the list would make it too large and too unrealistic to work with in only three years of ESL classes at the secondary level.

Limitations

Although the findings of the study and the resulting SEL are expected to be useful to both materials designers and ESL students, there are some limitations to this study that should be considered. First, I consider limitations of the vocabulary-focused approach taken in this study. I am not aware of any ESL textbook series that is specifically designed to present and systematically recycle high frequency words (though a notable exception is Schmitt and Schmitt's (2005) *Focus on Vocabulary: Mastering the Academ*ic *Word List*). Writers of integrated skills textbooks like *Quest* and other secondary ESL series designed for Quebec schools appear to have priorities other than vocabulary. These include motivating themes, authentic texts, tasks that promote oral interaction, and supplementary grammar support. Nowhere in the back covers or in the online promotional materials about *Quest* is there mention of a frequency-informed approach to vocabulary. Therefore, it is somewhat unfair to fault the materials for something they did not set out to do. I also acknowledge that taking into consideration the varied and complex structure of a textbook, using a list of frequent words to design it can be a very complicated task. I recognize that it may be challenging for textbook writers to include high-frequency vocabulary and, at the same time, make useful materials (e.g. in terms of grammar, thematic lexis, interesting activities, etc.).

In addition, the frequency-focused approach taken in this study assumes that all learners seek to reach a high level of proficiency in English and the emphasis has been on preparing learners for future academic study. More particularly, the research assumes that a high level of literacy in English is a goal shared by all. While I remain convinced that

knowing high frequency vocabulary is an efficient route to L2 literacy and to L2 proficiency generally, not all learners will see this as a personal aim. I recognize that words that are relevant to one student may not be of interest to all secondary students. Thus even though *aircraft* and *vehicle* are high frequency 2k words, some learners may never have any need to use them. Therefore, the words *plane* and *car* would meet their needs.

Another limitation is the narrowness of the study's focus, which is mainly on high-frequency families and their occurrence in the *Quest* materials. The study does not solve the issue related to practicality of teaching the mid-frequency words. It would be interesting in a future study to look more closely at the occurrence of mid-frequency words in a larger textbook corpus and the words that are missing. It seems likely that many mid-frequency words are French-English cognates that French-speaking learners of English can get 'for free' via activation of cognate awareness. Investigating this intriguing possibility was beyond the scope of this study.

There are also limitations related to the corpus methodology. The first of these pertains to the size of the corpus. Only one series of ESL textbooks was analysed in this study. Ideally, several other comparable series would be scanned and analyzed. This would help us understand whether *Quest* is typical or perhaps unusually rich (or poor) in its presentation and recycling of frequent vocabulary. A larger textbook corpus would also be useful in producing a more valid (and useful) pedagogical list. Clearly, the SEC cannot claim to represent the exact textbook vocabulary input received by all students in secondary ESL classes in Quebec. In order to obtain this figure, it would be necessary to

analyse all textbooks and extra materials used by all ESL secondary teachers in the area, which could be a very complex task. Also, in the absence of classroom data, we do not know the extent to which particular words were recycled in activities. It is also possible that teachers taught some of the words explicitly.

Finally, another limitation is the reliance on frequency lists based on the British National Corpus. This is not a serious shortcoming: Analyses using frequency lists based on British corpora consistently provide comparable frequency profiles of texts from American, Australian, Canadian and British sources (Martini, 2011, unpublished paper). Nonetheless, lists based on a Canadian corpus would have been ideal for this project. Unfortunately, such frequency lists and analysis tools do not yet exist for Canadian English.

Implications for Research

The analysis of the *Quest* series makes a research contribution by doing a textbook analysis that has not been done in Quebec before. The secondary ESL materials approved by the Ministry of Education in Québec have never been analysed for their vocabulary content and its implications in second language vocabulary acquisition. As a consequence, there was no clear description of the vocabulary input supplied by these books, leaving teachers with no guidance in terms of vocabulary instruction. The future direction for this research is to compile a more comprehensive corpus that includes all textbooks used in these secondary schools. This larger corpus would allow teachers to make more efficient choices for their Ministry-approved and supplemental materials.

In order to really understand the input the learners are exposed and the vocabulary in it, we also need studies that are more closely focused on language classrooms. An ideal study should include the different types and sources of input students are exposed to in the classroom. For example, we could combine analyses of the following sources: the current SEC; the information about which parts of the book students actually use in class or at home; the supplemental materials used in class; and finally, the recordings of teacher talk. All these sources of input could be analysed using the same tools I used to analyze the SEC. The results would show us a more comprehensive picture of the classroom input secondary students in Québec are exposed to. The current study is only the basis, the first step along the way to a much broader research goal.

Finally, this study can also be used in research that measures learner vocabulary knowledge at various frequency levels. To what extent do they know the words that are in the books? To what extent does the use of the workbooks improve their vocabulary knowledge? The SEC could be used to make a vocabulary test that allows teachers to assess students' vocabulary size before and after instruction. Consequently, it would allow teachers and materials designers to evaluate the effectiveness of the exposure and recycling of words in these books. It is not enough to know which words occur in the materials; we also need to know whether learners are picking up any of this language. Teachers would also have a better idea of how to guide the next stages of their vocabulary instruction.

Implications for Teaching

The current study shows that teachers would do well to use the entire content of the *Quest* textbooks in order to maximize exposure and take full advantage of the vocabulary that occurs in the materials. The books show a very good representation of the first 1,000 most frequent words in English, and some deficits at the 2,000 and 3,000 levels. The issue I found in a teacher survey done in secondary schools in Québec (Horst, et al. 2011) was that teachers do not use the entire content of the textbooks and do not select supplemental materials with the high-frequency vocabulary in mind. Taking materials from different sources can provide students with very interesting themes, but it is unlikely to result in systematic exposure to the vocabulary they need to know.

Teachers should be aware of frequent vocabulary and its importance for reading comprehension of texts in English. They should have access to frequency lists and make sure that they are giving attention to all of the 1k, 2k, and 3k level words. Once the students acquire these important words, they will be able to increase their vocabulary levels focusing on the important mid-level words (between the 3k and 9k levels). The current vocabulary knowledge of secondary students in Québec suggests that they do not master all the words up to the 3 k level, and that these words should be emphasised in their ESL syllabus.

How should teachers deal with the high frequency vocabulary? Teachers should teach vocabulary explicitly. As previously mentioned, White *et al.* (2012) showed that secondary students in Québec French-medium schools don't know all of the 1k, 2k and 3k words in English. Even if the books were able to expose them to many high-frequency

words and recycle them, exposure alone is probably not enough. Besides, there is very little focus on vocabulary in the *Quest* series. The authors give lists of words with their meanings, but they do not provide students with any other strategies to learn these words. In order to effectively teach these essential words, explicit vocabulary instruction should be part of their ESL classes.

Finally, teachers also have to deal with the mid-frequency level words. But how can they fit all these words in their classes considering their time and syllabus limitations? It is unrealistic to believe that teachers will be able to teach all these words in the three last years of secondary schools; however, it is reasonable to suggest that they choose certain words to be taught explicitly and that they provide students with strategies to improve their vocabulary levels. I suggest the three following steps for vocabulary selection and instruction using the SEL and/or the mid-level words. This is just an example of the many different vocabulary strategies teachers can use in their classrooms.

Step 1. In order to prepare students for their future academic studies in CEGEP and/or university, it is important that teachers select the words that are used in their academic settings. As an example, teachers could select words such as, *'accommodate'* and *'sufficient',* and include them in their supplemental materials (e.g. authentic texts could be adapted to include these words).

Step 2. Raise awareness of French-English cognates in class. Researchers recognize the importance and usefulness of cognates in teaching vocabulary, but they agree that these connections are not easily recognised by learners (Horst, Cobb, White & Martini, 2012; Moss, 1992; Otwinowska-Kasztelanic, 2009). Furthermore, about two-thirds of words

beyond the 2k level have French, Latin or Greek origins (Nation, 1990). If teachers focus part of their instruction time developing English-French cognate awareness, students will possibly be able to recognize a large number of words with little effort. As a result, they could benefit from their first language to improve their vocabulary size considerably. The examples in step one are English-French cognates, '*accommodate'* (EN) - *'accommoder'* (FR) *and 'sufficient'* (EN) – *suffisant* (FR).

Step 3. After working on the meaning of these words, students could be taught how to recognize and understand the inflected and derived forms of the same word. Teachers could use activities that facilitate the use of these forms and thereby increase their students' exposure to the vocabulary that is being used in the activities. For the examples cited above, the following forms could be emphasised:

Accommodate *=> accommodated, accommodates, accommodating, accommodation, accommodations*

Sufficient *=> sufficiency, insufficiency, insufficient, insufficiently, sufficiently*

Conclusion

There is increasing research interest in textbook corpora and classroom input (Alsaif & Milton 2012) with the aim of knowing more about the quality and quantity of classroom input learners are exposed to. The current study contributes to this line of research. In order to improve secondary Quebec learners' knowledge of vocabulary in English, it is important to understand and evaluate the pedagogical materials these students use in their ESL classes. These materials are, in most cases, the main or only

contact these learners have with their second language; for this reason, they should be systematically designed to attend learners' needs and to improve the acquisition of L2 vocabulary.

Although the scope of my research is limited, the analysis of the vocabulary content of Ministry-approved textbooks is unique in that it has not been done in Canada or Quebec before. It clearly shows deficits in the representation of important high-frequency vocabulary and the need for an investigation of a larger textbook corpus. Another strength of my study is that it explores a sequence of books, and the opportunities to develop vocabulary knowledge over time. If vocabulary acquisition is incremental, book series should consider their lexical content as a unit and develop it systematically throughout the books. There are great possibilities of improving vocabulary acquisition, but for this to happen we need instructional materials that prioritise high and mid-frequency word levels.

And finally, I hope that this study will be useful to teachers in a way that will help them use their existing Ministry-approved books more efficiently, and also help them with the selection of vocabulary for supplemental materials. I also hope that researchers and materials designers can works together to develop new priorities and solutions for teaching vocabulary. In order to improve the books available in the market, it is important to understand the deficits in students' L2 knowledge and to negotiate the realities of the classroom with the learner's needs.

References

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

Baxter, G., Beyea, C., & Ford, C. M. (2009). *Quest: English as a second language. Secondary cycle two, year three.* Montréal: Chenelière Éducation.

Beyea, C., Bougie, P., & Ford, C. M. (2008). *Quest: English as a second language. Secondary cycle two, year two.* Montréal: Chenelière Éducation.

Bougie, P., Capparelli, T., Ford, C. M., & Giannas, E. V. (2007). *Quest: English as a second language. Secondary cycle two, year one.* Montréal: Chenelière Éducation.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System, 32*, 251-263.

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes, 57*(2), 295-324.

Cobb, T. (2002). Review of Norbert Schmitt, *Vocabulary in Language Teaching.* – New York: Cambridge University Press. *Canadian Journal of Applied Linguistics.*

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning and Technology, 11*(3), 38-63.

Cook, V. (2001). Using the first language in the classroom. *Canadian Modern Language Review, 57*(3), 402-423.

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

Coxhead, A. J. (1998). *An academic word list* (English Language Institute Occasional Publication No. 18). Wellington, New Zealand: Victoria University of Wellington.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

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.

Granger S. (1996). Romance Words in English: from History to Pedagogy. In Svartvik J., (Eds). Words. Proceedings of an International Symposium, (pp. 105-121). Stockholm: Almqvist and Wiksell International.

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

Holley, F. M. (1973). A study of vocabulary learning in context: the effect of new-word density in German reading materials. *Foreign Language Annals, 6*, 339-347.

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics, 9*(1), 21-44.

Horst, M. (2009). Revisiting classrooms as lexical environments. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners: Papers and perspectives in honour of Paul Meara* (pp. 53-66). Clevedon: Multilingual

Matters.

Horst, M. (2010). How well does teacher talk support incidental vocabulary acquisition? *Reading in a Foreign Language, 22*(1)*,* 161-180.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207-223.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning & Technology, 9*(2), 90-110.

Horst, M., Cobb, T., White, J., & Martini, J. (July, 2012). *Towards a useful measure of French-English cognate awareness.* Paper presented at the 11th International Conference of the Association for Language Awareness (ALA), Montreal.

Horst, M., White, J., & Cobb, T. (October, 2011). *How many English words do Quebec secondary students 'know' and are they the 'right' words?* Paper presented at the annual conference of La société pour la promotion de l'anglais, langue seconde, au Québec (SPEAQ), Montreal.

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

Kachroo, J. N. (1962). Report on an investigation into the teaching of vocabulary in the first year of English. *Bulletin of the Central Institute of English, 2*, 67-72.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud, & H. Béjoint, (Eds), *Vocabulary and applied linguistics*, (pp. 126-132). London: Macmillan Academic and Professional Ltd.

Lee, J. (2006). Subjunctive were and indicative was: A corpus analysis for English language teachers and textbook writers. *Language Teaching Research, 10*(1), 80-93.

Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in quebec. *TESOL Quarterly, 28*(3), 563-579.

Martini, J. (2001, unpublished paper). English-Portuguese cognates in a newspaper word list.

Matériel Didactique Approuvé Pour L'enseignement Secondaire: Ensembles Didactiques 2011-2012. Ministère de L'éducation, du Loisir et du Sport Québec. Bureau D'approbation du Matériel Didactique.

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

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

Moss, G. (1992). Cognate recognition: Its importance in the teaching of ESP reading courses to Spanish speakers. *English for Specific Purposes, 11*(2), 141-158.

Nation, I. S. P. (1990). Teaching and learning vocabulary. Boston: Heinle & Heinle.

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

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

Nation. I. S. P. (1990). *Teaching and learning vocabulary.* Boston: Heinle & Heinle Publishers.

Nation, I. S. P., & Laufer, B. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics, 16*, 307-322.

Nation, I. S. P., & Wang, M. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, *12*, 355–380.

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

Otwinowska-Kasztelanic, A. (2009). Raising awareness of cognate vocabulary as a strategy in teaching English to Polish adults. *Innovation in Language Learning and Teaching, 3*(2), 131-147.

Pimsleur, P. (1967). A memory schedule. *Modern Language Journal, 51,* 73-75.

Richards, J. (2001). *Curriculum Development in Language Teaching.* Cambridge: Cambridge University Press.

Römer, U. (2004). 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. 151–68). Amsterdam: John Benjamins.

Saragi, T., Nation, P., & Meister, G. (1978). Vocabulary learning and reading. *System, 6,* 72-80.

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

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

Schmitt, N. (2011, March). *Mid-frequency vocabulary: The great gap in vocabulary research and instruction.* Paper presented at the annual conference of the American Association for Applied Linguistics, Chicago.

Schmitt, D., & Schmitt, N. (2005). *Focus on vocabulary :mastering the academic word list*. White Plains, NY: Longman.

Shin, D., & Nation, P. (2007) Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, 1-10.

Shortall, T. (2007). The L2 syllabus: Corpus or contrivance? *Corpora, 2*(2), 157-185.

Sinclair, J. (1997). Corpus evidence in language description. In A. Wichmann, et al (Eds.), *Teaching and language corpora. (pp. 27-39)*. Harlow: Longman.

Statistics Canada. (2011). *Population by language spoken most often at home and age groups, 2006 counts, for Canada, provinces and territorie:20% sample data*. Retrieved from http://www12.statcan.ca/census-recensement/2006/dp-pd/hlt/97-555/T402-eng.cfm?Lang=E&T=402&GH=4&SC=1&S=99&O=A, December 7, 2011

West, M. (1953). A general service list of English words. London: Longman, Green.

White, J., Martini, J., Horst, M., Cobb, T. (May, 2012). *Towards a corpus informed vocabulary pedagogy for Quebec secondary ESL learners.* Paper presented at the annual meeting of the Canadian Association of Applied Linguistics (ACLA/CAAL), Congress of the Social Sciences and Humanities, Waterloo.

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

Wilkins, D. A. (1972). *Linguistics in language teaching.* Cambridge, MA: MIT 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://www.natcorp.ox.ac.uk/, 2012.

Cambridge English Corpus (CEC). In Cambridge English Language Teaching. Retrieved from http://www.cup.cam.ac.uk/fr/elt/catalogue/subject/custom/item3637700/?site_locale=fr_FR, 2012.

Corpus of Contemporary American English (COCA). Retrieved from http://corpus.byu.edu/coca/, 2012.

Corpus of English as a Lingua Franca in Academic Settings (ELFA). Retrieved from http://www.helsinki.fi/englanti/elfa/index.html, 2012.

Data Sets

Statistics Canada. (2012). Population by language spoken most often at home and age groups, 2006 counts, for Canada, provinces and territories - 20% sample data Retrieved from http://www.statcan.gc.ca.

Appendix A

The SEL: this list contains only the 1,281 headwords of each word family

abroad	annoy	bark	bond
absence	anonymous	barn	bone
abuse	anti	barrel	bonus
accelerate	apart	barrier	book
accent	apartment	basement	boom
accommodate	apology	bash	boost
according	appal	basis	border
account	apparent	basket	borough
accountant	appliance	bat	boss
accuse	appoint	bath	bottle
ache	appreciate	bathroom	bounce
acid	approve	battery	boundary
adequate	approximate	beef	bow
adopt	arrange	beer	bowl
advise	arrears	beforehand	box
aerial	arrow	beg	bracket
aeroplane	articulate	behalf	brake
affair	artificial	belt	branch
affection	ashamed	bench	brass
agenda	asleep	bet	breakdown
aggressive	assemble	bible	breast
aircraft	assess	bid	breed
airline	assume	bin	brick
airport	assure	bind	bridge
album	astonish	bingo	brilliant
alcohol	atmosphere	biscuit	brochure
alert	authorise	bishop	brush
alive	authority	bite	buck
alley	automatic	bitter	bucket
allocate	autumn	blade	bug
almighty	avenue	blame	bulk
alongside	average	blank	bull
alright	awful	blanket	bump
alter	awkward	blind	bunch
alternative	backside	block	burden
altogether	badge	bloom	burgle
aluminum	bake	boil	burst
ambulance	banana	bold	bush
anger	bang	bolt	bust
anniversary	bargain	bomb	butcher

butter
button
buzz
bypass
cabinet
cable
cake
calendar
camel
campaign
can
canal
cancel
candidate
candle
canteen
capacity
capital
captain
capture
cardboard
carol
carpet
cart
cash
cassette
castle
casual
casualty
catalogue
cater
cathedral
catholic
ceiling
cement
certificate
chairman
chamber
champion
channel
chapel
charity
charm
chat
cheat
cheek
cheer
cheese
chemist
cheque
chest
chimney
chip
choke
chop
chuck
chunk
cigarette
cinema
circuit
civil
civilise
clap
clash
clerk
clever
click
cliff
clinic
clip
cloth
cloud
clutch
coach
coal
coat
cock
coin
coincide
collapse
collar
colleague
column
comedy
command
commission
commit
compact
compensate
compile
complaints
complicate
compliment
comprehensive
compromise
conceive
conclude
condense
confess
confidence
confidential
conflict
congratulate
congregate
conscious
consequent
conservative
considerable
consistent
continent
convenience
convert
cop
cord
core
corridor
cottage
cotton
cough
counter
countryside
county
court
cracker
craft
cramp
crap
crawl
creep
cricket
criminal
cripple
crisis
crisp
criterion
criticism
crown
crucial
cruel
crush
crystal
cube
cupboard
cure
curious
current
cushion
custom
cylinder
damn
damp
dash
data
deaf
debt
decent
declare
deduct
deed
defence
delegate
deliberate
delicate
delight
deliver
democrat
deny
department
depress
deputy
derby
desert
deserve
detach
deteriorate
devil
devote
dial
dialect
diary
dictate
digest
dimension
dine

dinosaur
dip
directory
disaster
discipline
discount
discriminate
disgrace
disguise
disgust
display
distinct
distress
distribute
district
ditch
dive
divorce
dock
document
dodgy
dole
domestic
dominate
doorstep
doorway
dose
dot
dozen
drain
draught
drawer
dread
dreadful
drift
drill
drip
drown
dual
duck
duke
dull
dummy
dump
dust
duty
dye
ease
east
echo
egg
elbow
elder
eldest
elect
eleven
embarrass
embassy
emergency
empire
enable
engage
enquire
enthusiastic
entitle
envelope
equivalent
escort
essay
estate
evidence
evolve
ex
exaggerate
excess
exclude
exhibit
exhibitionist
exit
expand
experiment
explode
export
expose
extend
extract
facility
factor
fail
fame
fancy
fare
fat
fatal
fault
fax
feather
fee
feed
felled
fence
fertile
fetch
fiddle
fierce
filter
fine
fir
firm
flag
flame
flap
flare
flat
flatting
flavour
flexible
flight
float
flood
flu
fold
folder
fond
ford
forecast
foreign
foresee
fork
forth
fortnight
foster
fraction
framework
fringe
frost
fry
fume
funeral
fur
furnish
furniture
furthermore
fuss
gallery
gallon
gamble
gang
gap
garage
garden
gate
gee
generous
gentleman
genuine
gift
glad
glance
glaze
globe
glory
glove
glow
golf
goodbye
gospel
grab
grace
gracious
gradual
grain
gram
grant
graph
grateful
greed
grief
grind
grip
gross
guarantee
guilty
gut

hairdresser
ham
hammer
handicap
handy
harass
hardware
harry
harvest
hassle
hay
headache
headline
headmaster
heal
heather
heaven
heck
hedge
heel
height
helicopter
hell
hen
hesitate
hint
hobby
hole
holy
honey
hood
hook
hooray
horrendous
host
housewife
hut
ideal
identical
idiot
idle
immense
implicate
impose
incentive
inch
incidental
incline
indeed
index
indulge
infect
inherit
initiate
inject
inn
insist
inspect
install
instance
insult
insure
intellectual
inter
interfere
interpret
interval
intrigue
iron
irony
irritate
isle
jack
jacket
jam
jar
jaw
jeans
jewel
jog
joint
jolly
jug
juice
junction
junior
justice
keen
kettle
kidding
kidney
kit
knife
knot
know
ladder
lamb
lamp
landlord
lane
lap
lawn
layer
layout
lazy
leaf
leaflet
league
leak
lean
leap
lecture
leisure
lend
length
lever
liable
liberal
licence
lid
likeness
liquid
literal
litre
litter
loan
lock
lodge
loft
log
loop
loose
lord
lump
lunchtime
lung
luxury
madam
magnet
manoeuvre
manor
manual
mar
march
margin
marvel
mask
massive
master
mate
mature
maximum
mayor
meantime
mechanic
medicine
melt
mend
menu
merchant
mere
merit
merry
micro
microphone
mid
mild
military
millimetre
miner
mini
minimum
minister
minor
minus
miracle
mirror
misery
mission
mix
moan
mobile
mock
monitor

monopoly
moor
moral
mortal
mortgage
motion
motor
motorbike
motorway
mouth
mud
mug
multi
multiple
multiply
nail
naive
naked
narrow
nasty
naughty
nay
neck
needle
neglect
negotiate
nerve
nest
neutral
nevertheless
niece
nightmare
nip
non
none
nonsense
norm
north
nuclear
nuisance
numerous
nursery
nut
oak
object
obscure
obtain
occupation
occupy
offence
offend
onwards
oral
orchestra
organ
orient
origin
otherwise
outcome
outrage
oven
overall
overcome
overhead
overnight
overtake
overtime
owe
pace
pack
pad
pain
pal
palm
pan
panel
panic
paperwork
par
parade
parcel
parliament
passage
passenger
pat
patch
pathetic
pause
pave
peak
peculiar
pen
penalty
pencil
penny
pension
per
perception
phenomenon
photocopy
pie
pier
pig
pigeon
pilot
pin
pinch
pint
pipe
pit
pitch
pity
plaster
plate
platform
pleasure
plenty
plough
plug
plumb
plus
poison
poke
pole
polish
poll
pond
population
port
portion
pose
posh
postcard
pot
potato
pottery
pound
pour
powder
praise
pre
preach
proceed
precise
pregnant
premise
prescribe
presume
previous
pride
prime
prince
princess
principle
privilege
pro
proceed
profit
prominent
prompt
pronounce
proof
proportion
prospect
pub
pudding
pump
punch
punish
pupil
purchase
pure
purple
purse
puzzle
qualify
quantity
quarter
query
questionnaire
rabbit
race
rack
rag

rail
random
rank
rarefy
rat
rattle
raw
ray
rear
receipt
reception
recession
recipe
reckon
recommend
recruit
redundant
referee
refresh
refrigerate
register
regret
regulate
rehearse
relevant
relief
relieve
religion
remote
repair
reproduce
reputation
rescue
resign
resist
resort
restore
restrict
retail
retire
reverse
revolt
revolution
rhyme
ribbon
rid
ridiculous
rip
rob
rocket
roof
root
rope
roses
rot
rotate
rough
route
rove
row
royal
rub
rubber
rubbish
rude
rugby
ruin
rumour
sack
sacrifice
saint
sake
salt
salvation
sample
satisfy
sausage
scale
scheme
scope
scout
scrap
scrape
scratch
screw
scribble
scrub
seal
second
secretary
seed
seldom
semi
senior
sensible
sensitive
series
session
severe
sew
sex
shadow
shame
shave
shed
sheer
shelf
shell
shelter
shield
shift
shirt
shore
shove
shovel
shy
signature
silk
silly
sin
sincere
skeleton
skip
slap
slash
slave
sleeve
slice
slim
slip
slot
smack
smash
smooth
snap
sneak
sniff
soak
socialism
sock
sole
solicitor
solid
somewhat
sophisticated
sore
soul
soup
south
southeast
southwest
spare
spark
spectacle
speculate
spit
spite
splash
splendid
split
spoil
sponsor
spoon
spray
squash
squeeze
stab
stack
stain
stairs
stake
stamp
starve
status
steady
steam
steel
steep
steer
stick
stiff
stir
stitch
stock

stomach
stone
straightforward
strain
strap
straw
strict
stride
string
stripe
stroke
struggle
stupid
submarine
subsidy
substance
subtle
suck
sue
suffer
sufficient
sugar
suicide
suite
sum
super
superb
supermarket
supper
supplement
surgeon
surgery
surname
suspend
swallow
swan
swap
sweep
swing
switch
sympathy
symptom
tab
tablet
tack
tackle
tag
tail
tank
tape
tax
taxi
tea
tease
technical
telecom
temper
temperature
temporary
tempt
tender
tennis
tent
terminal
terminate
terrace
terrific
text
theory
therapy
thirteen
thirty
thorough
thrill
throat
thumb
thus
tide
tidy
tie
tight
tile
tilled
timber
timetable
tin
tissue
toe
toilet
token
tokenism
tongue
torch
toss
tough
towel
tower
toy
trace
traffic
tragedy
tragic
translate
trap
tread
tremendous
trial
triangle
trivial
trolley
troop
trophy
trousers
tube
tumble
tunnel
turkey
twelve
twin
un
underground
underline
undo
uniform
unless
update
upper
upwards
urge
urgent
utilise
utility
vacuum
vague
valid
van
vandal
various
vast
vat
veggie
verbal
verse
versus
vest
vet
via
vice
vicious
victim
victory
village
virgin
virtual
virus
visible
vital
volume
voluntary
voucher
wagon
wallet
wallpaper
ward
wardrobe
warehouse
warrant
weak
weapon
wed
weed
weight
weird
welfare
whack
whatsoever
whereabouts
whereas
whereby
whip
whiskey
whistle
whoop
wicked

widow	wire	wreck	zero
wig	withdraw	wrestle	zone
wine	wobble	wrist	
wing	wool	yard	
wipe	worth	yellow	

Appendix B

This is a sample of the SEL: The complete word family for each headword is shown. Interested readers can contact Juliane Martini (jupisani@yahoo.com) to obtain the full list.

adequate
- adequacy
- adequately
- inadequacies
- inadequacy
- inadequate
- inadequately

adopt
- adopted
- adopting
- adoption
- adoptions
- adopts
- adoptive
- adopter
- adopters

advise
- advising
- advises
- advisers
- adviser
- advised
- advisable
- advisably
- inadvisable
- advisor
- advisors
- advisory

affection
- affections
- affectionate
- affectionately
- affective
- affectively

aggressive
- aggressively
- aggressiveness
- aggressivenesses

allocate
- allocated
- allocates
- allocating
- allocation
- allocations

alter
- alterable
- alteration
- alterations
- altered
- altering
- alters
- unalterable
- unaltered

annoy
- annoys
- annoyed
- annoying
- annoyance
- annoyances

apology
- apologize
- apologies
- apologetic
- apologetically
- apologist
- apologists
- apologizes
- apologized
- apologise
- apologising
- apologizing
- apologises
- apologised

appoint
- appoints
- appointed
- appointing
- appointments
- appointment
- reappoint
- reappointed
- reappointing
- reappointments
- reappointment
- reappoints

appreciate
- appreciable
- appreciably
- appreciated
- appreciates
- appreciating
- appreciation
- unappreciated
- appreciative
- unappreciative
- appreciatively