2. Review of Literature

2.1 Overview

In order to understand topics related to this research on vocabulary coverage, one must first understand vocabulary learning in general as well as corpus linguistics. This literature review will give general descriptions of such applicable topics to the current study. The first section of the literature review (2.2) deals with research in the general area of vocabulary learning. This section explores different theories on language teaching and how vocabulary teaching is approached (2.2.1) and the affects that age might play on second language learning (2.2.2). This is followed by a section that describes the differences between second and foreign language learning (2.2.3) then beliefs of how vocabulary should be ideally learned (2.2.4). These topics of discussion give further insight into the motivation for and bases of, or theoretical beliefs behind, current studies like this one.

The second half of this review is more specific, contextualizing research in corpus linguistics. It first gives historical perspectives about corpora and their use in applied linguistics (2.3.1). From there, through this discussion, more specific areas of related research are described and then compared with each other and also with the topics germane to the current study in question. The issues addressed include native speaker intuition (2.3.2), lexical frequency (2.3.3), vocabulary size (2.3.4), and word lists (2.3.5). The review then draws to a close with a discussion of implications for pedagogical material analysis and development. In this final section (2.3.6), the replicated study is described, and the applicability of its findings to the current research is discussed.

2.2 Vocabulary Learning

2.2.1 Theoretical Perspectives and Approaches

Trends in vocabulary learning in SLA can be described in terms of the histories of the theories and approaches to language learning in general. According to Bade (in press), from the time of the ancient Greeks more than two and a half millennia ago until the twentieth century, structural methods of teaching second languages were predominant in Western cultures (p. 146). Because of views of language as being basically grammatical patterns, communicative competence and vocabulary were often neglected. In the nineteenth century, an approach based on these beliefs was developed and became known as the Grammar Translation approach. According to Richards and Rogers (2001), the Grammar Translation approach was widely used from the mid-nineteenth to the mid-twentieth centuries when Western students were generally taught Latin, a dead but written language (pp. 5-7). They describe this approach's stance on vocabulary as word-forms being mere pieces of sets of rules used to create translation equivalents on the sentential level. Students were taught a rule, given a list of words and their translations, and then asked to translate sentences to and from the target language. Because of this, there was not much of a connection to communication or even to the meaningfulness of a sentence. The modality of language was almost always written, and there was little or no context for the reader to determine overall meaningfulness. For example, a sentence written or read by a student could be grammatically correct, but the words in that sentence can cause it to not make sense or for it not to be socially appropriate.

Nearing the end of the Grammar Translation approach's prominence, another method of language teaching and learning was developed based on the beliefs that natural communication was the best way to learn a second or foreign language. According to

Richards and Rogers (2001), towards the end of the nineteenth century, the Direct Method was developed (p. 11). This approach to language teaching was based on the belief that the teacher and materials for second or foreign language teaching should only address the students in the target language. In other words, as Richards and Rogers (2001) describe, only the language being learned is used, and there would be no translations given or assigned (pp. 11-14). This is particularly important in terms of second language learning and teaching because this method may simulate the processes a person faces in a non-native environment: how he or she would have to find ways to communicate and learn in such an environment, possibly not having any interlocutors who speak his or her native language. The input that a student receives in this method is in the form of communicative sentences in which the meanings or functions of the words generally have to be induced from context. In terms of vocabulary, besides induction from context, words have to be taught using tangible objects or motions or through association with other words learned in the target language. According to these authors, with such a complex method of matching a word-form to a concept, basic vocabulary needed for everyday communication is emphasized for efficiency reasons. Because large amounts of time and energy are required for their instruction, these frequent or useful vocabulary items are chosen carefully as not to waste time on items not likely to be encountered or used.

During the second half of the twentieth century, another new method for language teaching was developed and came into prominence. This method was labeled the Audiolingual Method. This method's approach was based on structural linguistics and behaviorism. Richards and Rogers (2001) describe how proponents of this method believe that language is a set of rules from parts as small as sounds all the way up to the sentential level (p. 54). These rules, according to this approach, can be taught best through repetitive exercises like drills. Behavioral psychologists, Richards and Rogers write, believe that the way a person learns (including the learning of a language) is by receiving a stimulus, making a response to that stimulus, and then receiving positive or negative reinforcement (p. 56). In terms of language learning, this means that a student is given input in the target language and then asked to answer or repeat the stimulus. If the student answers correctly, he or she is given positive reinforcement and learns that such behavior (i.e. giving the correct answer) is good. If he or she answers incorrectly, however, negative reinforcement is applied, and such mistakes are shown to be bad, hopefully making the student try harder the next time so as not to receive punishment. Dialogues and drills making use of repetition and memorization are used in this method as the majority of activities. These activities involve small contexts of culturally-based language as single sentences or multiple-line conversations. In terms of vocabulary acquisition, it is from these pieces of language context that students are supposed to inductively learn the meanings to individual words. Word meaning, proponents claim, cannot be learned in isolation (Richards and Rogers, 2001, p. 64), so any vocabulary instruction essentially would take place through induction from meaningful contexts.

The most recent and widely accepted methods to language teaching make use of the Communicative Approach. This approach is so named because of the importance it places on language being a means of communication instead of, for example, a set of rules. In the 1970s and 1980s, particularly, more communicative-based approaches began to gain popularity in language teaching. As transportation and technology advanced, so did the economic interdependence of European countries (Richards and Rodgers, 2001, p. 154). People needed to be able to communicate in real time with each other in various types of interpersonal situations. Thus, the goal of language learning was no longer to memorize grammatical rules but to achieve real communicative competence. Language teaching theories similarly changed. Summarized by Richards and Rogers, an important similarity across different communicative-based approaches is that they generally emphasize real communication through activities, task performance, and contextually meaningful and if possible authentic language use (p. 166). The usual objective of using the Communicative Approach is for the student to be able to communicate with speakers of the target language, so it is seen as useful to use authentic texts in order to give students an idea of how the language is used in the "real world."

One of the more prominent of these branches of the Communicative Approache is known as the Natural Approach, developed by Terrell and Krashen (1983). Its development also brought more specific ideas on how languages could ideally be learned. Ideally, Richards and Rogers (2001) summarize, in the Natural Approach, a target language should be acquired instead of learned. *Learning*, according to Terrell and Krashen, refers to one being taught and consciously developing an understanding of what it is that needs to be learned. *Acquisition*, in Terrell and Krashen's view, naturally occurs in a more subconscious manner without effort, similar to how children acquire their first language. One of the characteristics of this method is that the student receives a large amount of input, much in the same way that a child acquiring his or her first language does. Although there is a lot of input, second language acquisition is believed to occur best when a person is exposed to input in a structured way. According to Terrell and Krashen, the input should continuously be altered to be slightly more complex and to include slightly more previously unknown items than their level of competence at the time (p. 32). In terms of vocabulary, Richards and Rogers also explain how in the Natural Approach, the lexicon is given importance as a means of creating and understanding meaning. New vocabulary is expected to be acquired through induction, using context or visual cues. Translation or use of the students' first language is not desirable in this approach. However, as described in the research of Bley-Vroman (1989) (see section 2.2.2), there are serious doubts to the similarities between a child's first language acquisition and an adult's second language learner.

Another advance in the area of language teaching theory that is not necessarily part of the communicative-based approaches is one in which lexical units are the center of language learning and teaching. Following the work by Sinclair and Renouf (1988), Willis (1990) developed *The Lexical Syllabus*, which is based on teaching frequent word forms in the target language, English. In this plan of study, not only are frequent word-forms given attention, but frequent patterns and collocations (combinations) of words are also given importance. According to Richards and Rogers (2001), these lexical approaches, like *The Lexical Approach* developed by Lewis (1993), reflect a belief that the lexicon is central to both language and communication. Of particular importance, according to these beliefs, are frequent phrases or "chunks." These frequent clusters of words are seen as lexical units that should ideally be learned together instead of as parts of a whole. For example, in English, the phrasal verbs *put up with*, *put on*, *put away*, *put together*, *put out*, etc. all have different conceptual meanings that do not, necessarily, have much to do with the core meaning that the verb *to put* has when said in isolation. Richards and Rogers explain how these lexical approaches are not necessarily full approaches, but are more of ideas that could be applied to various existing approaches (p. 138). Particularly because methods based on the Communicative Approach continue to be dominant in language teaching, such emphasis on vocabulary may be added to a syllabus in a supplementary way. Similarly, a communicative-based syllabus can also supplement its teaching with some explicit teaching of grammar rules and translations. These additions, making the overall approach more eclectic, do not take away from the central goal of communicative competence.

2.2.2 Critical and Fundamental Difference Hypotheses

With the discussion of the Direct Method and other Communicative Approach methods, it should be made clear, however, that a post-pubescent (adult) student in a program does not necessarily learn the same way as a pre-pubescent (child) learns his or her native language as some proponents of these methods have claimed. Instead, as Bley-Vroman (1989) describes, second language learning by adults is fundamentally different from the native language acquisition of children (p. 49). Bley-Vroman labels this as the ‘Fundamental Difference Hypothesis,’ and based this on the Critical Period Hypothesis which, according to Griffiths (in press) believes that first language acquisition is run by universal grammar (UG) mental processes, but after puberty such devices no longer function as they do at birth (p. 35). Bley-Vroman claims that the second language acquisition of adults, unlike children, is driven by general problem-solving cognition (pp. 50-62). Hall (2005) presents some examples of such differences in second language acquisition. These include the presence of a first language from which to relate the target language, the full cognition and socialization of the learner, the necessity of instruction, as well as other variables that are seen as constants in first language acquisition (p. 234). Discussion of age is particularly relevant to the current study because the materials being studied were designed for adult learners; who, according to Bley-Vroman (1989), learn language in a fundamentally different way than children. Age is relevant to this study and others like it because they use corpora that reflect fully competent speakers of a language. Frequencies of lemmas are only relevant if the speakers or learners understand the concept that such forms represent. For example, an adult learner of Spanish might learn the word *grado* [grade, degree]. Such a feat would require only the mapping of a new form to an existing concept in his or her mind, which probably already has a lexical assignment in the first language. Children on the other hand have yet to gain much conceptual knowledge common in adults, so they should not be expected to learn frequent entries simply because they are common in adult speech.

2.2.3 Second vs. Foreign Language Learning

Not all language learning is the same. Besides the age of acquisition, another way to distinguish the type of language learning taking place is by the environment in which the language is being learned. As described by Cook (2003), a *foreign language learner* is one who is learning a language that is not a socially necessary language to speak in his or her own immediate cultural context. On the other hand, a learner could be living in a foreign culture where the people speak a different language than his or her own. When this is the case, the student learning the language of that new culture is said to be a *second language learner*.

In terms of vocabulary, Nation and Waring (1997) describe a key difference in motivation for controlled and optimized vocabulary learning in a second language compared to that of a foreign language:

> Teachers of ESL may be interested in measures of native speakers' vocabulary size because these can provide some indication of the size of the learning task facing second language learners, particularly those who need to study and work alongside native speakers. (p. 7)

This describes a particularly urgent need for learners in a second language environment to have communicative competence. It may not affect the everyday life of a foreign language learner to learn relatively infrequent forms at the expense of frequent or useful ones. If a foreign language learner, for example, learns infrequent lexical items at the expense of more frequent items, his or her daily life outside of the classroom would probably not be affected. Similarly, compared to a second language learner, a foreign language learner does not have a communicative need to be able to produce and understand his or her non-native language inside or outside of the classroom. A second language learner, however, may not have the luxury of being able to communicate in his or her native language outside of the classroom. This added necessity for communication and available environments for practice would probably help a second language learner learn more vocabulary at a faster rate than his or her foreign language learner peers. According to Nation and Waring (1997), when post-pubescent learners are in a second language environment, their rate of vocabulary growth in the target language is so significant that it is similar to the rate of vocabulary growth in adolescents in their first language (p. 8). In other words, on average, a student who studies a language abroad is able to, with enough motivation and the appropriate attitude, increase his or her target language vocabulary at similar rates as an adolescent learning vocabulary in his or her first language.

Because second language learners are in the environment of and surrounded by native speakers of the target language outside of the classroom, there might also be significant problems of errors or miscommunication. A person living in an environment where his or her language is not spoken may have difficulties in business transactions and/or social relationships outside of the classroom because of an inability to competently produce or sufficiently understand his or her nonacademic interlocutors. For example, Day, Chenoweth, Chun, and Luppescu (1983) investigated error corrections of the target language offered by native speakers to their non-native speaker interlocutors. The researchers categorized these error corrections, finding that vocabulary errors were the largest category of corrections, accounting for more than twice the amount of corrections based on syntactic errors. This shows the relative social importance of vocabulary in a second language environment compared to other aspects of language such as grammaticality and pronunciation. This could be because words hold important conceptual meaning while simple grammatical or allophonic mistakes may not have as much of an affect on the meaning of the message.

2.2.4 Decontextualization and Explicit Teaching

Another aspect of vocabulary learning is the manner in which vocabulary items are taught and learned. Ellis (1994) describes the benefits for implicit and explicit methods of vocabulary acquisition. He describes how vocabulary acquisition is generally implicit, citing evidence from studies on children rapidly acquiring the vocabulary of a first language and amnesiacs with damaged explicit memory abilities who are still able to implicitly learn (p. 268). However, Ellis also concedes that cognitive mediation is required to connect form with meaning, and that this form of conceptualizing one’s input relies on explicit learning (p. 268). This metacognition is seen as a form of explicit learning because the learner is actively conceptualizing while processing the input he or she receives.

Not all second and foreign languages are taught with a balance between implicit and explicit learning. Sökmen (1997) writes about how many language professionals are heavily influenced by naturalistic and communicative beliefs about language learning, which emphasize, “implicit, incidental learning” (p. 237). However, Nation (2001) makes the claim that “learners need to focus on words not only as part of the message but as words themselves” (p. 199). He describes how noticing is the first step in the learning process. To begin to learn or remember a word and its use, the learner must first notice its presence and significance. This is the basic idea that noticing is required for learning, and as defined by Krashen (1985), which one’s input does not always translate into one’s intake. Sökmen (1997) goes further and describes why strictly implicit instruction of vocabulary is not ideal. Learners, she claims, are not likely to guess correct meanings from written context and their comprehension of written texts, as a whole, is low when words are not previously known. Along with these downsides to implicit learning, Sökmen also found that guessing from context, even when correct, does not necessarily convert to long-term memory (p. 238).

Nation (2001) argues that explicit and decontextualized instruction of vocabulary should be used as a necessary supplement to contextual instruction through induction associated with widely used methods associated with the Communicative Approach (pp. 119-120). One such method used to explicitly teach vocabulary out of context is by giving the learner a definition. Brett, Rothlein, and Hurley (1996) found that there are significant benefits in vocabulary learning when students receive the definition of unfamiliar words as they occur in a story. By taking the word out of context, the instructor is showing his or her students that it is an item worth noticing, improving the chances that it will be learned and remembered. Other ways in which words can be decontextualized include pre-teaching, word-banks, glossaries, highlighting, and repeated encounters in a variety of contexts.

2.3 Corpora, Frequency, and Acquisition

2.3.1 Corpus Linguistics

Corpus linguistics refers to the study of corpora, which are large databanks of language that has actually occurred in real life from different genres. Leech, Rayson, and Wilson (2001) describe how since the late 1960s, linguists have been able to take advantage of computer processing to store and better understand language. With the advent of computers, a corpus could contain millions, then tens of millions, and more recently, hundreds of millions of words. One of the major purposes for such large databanks of real language is to better understand, as Sinclair (1991) describes, the *naturalness*, or textual well-formedness, of a given language. These corpora can even be designed to separate and mark dialect, recognize grammatical aspects of words, determine frequent word combinations (collocations), and measure relative frequencies of lexical entries' occurrences (Leech et al., 2001, pp. x-xi).

Corpus linguistic research is investigative or observational in nature. Unlike some other forms of linguistics where experiments are performed on participants, corpus linguistics relies on what has already been said or written, with goals of collecting, categorizing, and analyzing utterances in natural environments. This makes the discipline almost entirely observational or exploratory. The data are obtained through pre-existing utterances and are then explored. In other words, written and spoken texts are gathered from any number of sources (speeches, interviews, reports, textbooks, literature, etc.) then catalogued, creating a corpus. Sinclair (1991) further explains how corpus linguistics needs to continuously advance in order to not "misrepresent a language" and should never "offer as an instance of language in use, some combination of words which we cannot attest in usage" (p. 6). The more utterances collected and the better they are categorized, the more likely a language is to be well described. Such explanation of ‘natural,’ however, needs to be operationalized, especially because there is another branch of linguistics that uses the term. The naturalness described by Sinclair is not the same as that of a branch of computational linguistics called ‘natural language processing,’ in which computer programs are equipped with grammar rules and vocabulary lists create and interpret utterances. In Sinclair’s use of the term of naturalness, ‘natural language processing’ should be seen as artificial because it is not, necessarily, based on actual utterances. In order to make such computational fields more natural, however, it would behoove researchers to conduct further research in corpus linguistics to better describe natural language instances and processes.

One way that corpus linguistics relates to language teaching is that by studying how native-speakers use their own language, one can postulate ideal ways for non-native speakers to learn it. This is not to say that non-native speakers would be likely to learn the target language the same way that native speakers acquired it; corpus studies generally do not describe the process of acquisition but show how already competent speakers use the language. Instead, information gathered from such research could allow language planners to determine, for example, the general vocabulary needed for relative communicative competence based on vocabulary entries’ frequencies in the target language and possibly target regional variation. Because corpora can be coded for variations, modalities, and registers; different vocabulary may be determined as important, depending on the learners’ needs. Similarly, frequently occurring structures and collocations could be determined then given more or less emphasis in the teaching process. For an even more general example of the influence of corpus linguistics on second language teaching methodology, Cook (2003) writes that corpus linguistics has shown that native speakers tend to rely on chunks of language, possibly more than productive patterns, in speaking their native language. Because of this, some researchers and program directors have called for second language approaches to take some of the pedagogical emphasis away from grammar teaching and put more towards vocabulary and collocation teaching.

2.3.2 Intuition

One of the most obvious benefits derived from research in corpus linguistics is that it allows researchers to study linguistic occurrences (of words, collocations, structures, etc.) in real language. Stubbs (2001) writes that since the 1980s there has been a significant shift in applied linguistics from what Chomsky (1988) refers to as I-language (Internal, or of an individual) to that of E-language (External, or of a speech community). Thus, more importance has been placed on natural or real language use as a whole, as it is spoken and understood across a speech community compared to the internal language and introspection of that language by a single speaker. By using corpora, a researcher, textbook designer, teacher, or student does not need to rely on intuition or unsubstantiated beliefs about language to make claims of frequency or patterns of usage. Unfortunately, however, according to Biber and Reppen (2002), language-learning materials such as textbooks are usually only subject to intuitions of the authors and anecdotal (instead of empirical) evidence (pp. 205-206).

These intuitions and cultural beliefs regarding language that influence the design and word choice in textbooks do not always mirror reality. Sinclair and Renouf (1988) describe this inconsistency, explaining how the human mind is not designed to consciously recognize what is common or frequent in language (p. 151). Basic, highly frequent aspects of language are instead so commonplace that speakers do not take much notice of them.

Instead, what is noticed is that which differs from normal or frequently occurring uses of language. Hunston (2002) concedes that a speaker of a language can consciously know the relative frequency of some linguistic features, such as words, but only intuitively. For example, a native speaker of English probably could correctly choose *give* as being more common than *bequeath*, because *give*, according to Biber and Reppen (2002, p. 205) is one of the twelve most frequent lexical verbs in English, and *bequeath* may never have even been used by the given speaker.

However, not all lexical decisions are as intuitively clear as the example above, comparing a very highly frequent verb to a much less frequent verb. Between entries in adjacent frequency ranges, the ordering by frequency might be more difficult. Take, for example the following five professions of moderate frequency, of which according to Davies (2006) all are in the top in 2,000 Spanish lemmas, might not be as easy of a task (the frequency number is listed in parentheses):

dueño	[owner]	(1093)
soldado	[soldier]	(1568)
maestro	[teacher]	(961)
abogado	[lawyer]	(1680)
oficial	[oficial]	(1781)

Practically, in a section on professions and careers in a textbook, the author may ask himself or herself which professions the book should present. Experimentally, future psycholinguistic research could be combined with corpus linguistics to determine more precisely how well native-speakers of Spanish are able to determine relative frequencies.

Generally, textbook writers are language professionals and as such should view language empirically. Ideally, vocabulary decisions would be made based on empirical evidence of frequency and coverage across a target variation of the language. Biber and Reppen (2002) in an examination of ESL textbooks, however, found that this is not always the case. Before measuring the appropriateness of the vocabulary in these textbooks, the authors first studied corpora of English. They found that out of all the verbs in English, there are only 12 lexical verbs that occur more frequently than 0.01% (more than 1,000 instances per million words). From this, their motivation in measuring the appropriateness of the textbooks was to determine whether these twelve extremely frequent verbs were given particular attention. In this survey of 12 textbooks, the researchers found that 7 of these 12 most frequent lexical verbs were completely disregarded by all of the textbooks studied. This should give particular motivation for further study in the area of materials design as it relates to authentic production in the target language.

2.3.3 Frequency

For communicative competence (see section 2.2.2), there is obvious need for second language learners to be taught vocabulary that will be useful to them, especially because they are living in an environment in which their native language is not necessarily spoken. In general, one can assume that the most useful vocabulary would be those lexical items that are most frequently used by speakers of the target language. But, before discussing word frequency, first the term *word* needs to be operationalized.

As described by Sinclair (1991) as discussed in section 1.5, a *word* (orthographic word, or word-form) is a meaningful or functional group of connected letters, separated on either side by a space. In corpus linguistics, however, words are often described in terms of their lemmas. A *lemma* is a way to describe a group of word forms that are related by inflectional differences. In English, for example, Nation (2001) describes a *lemma* as a representation of a group of words, “consist[ing] of a headword and … its inflected and reduced [n’t] forms” (p. 7). Lemmas offer insight into second language acquisition because, according to Davies and Face (2006), once a learner is able to understand and produce the inflectional system, the individual, inflected word forms are relatively easy to understand and produce once one of the forms is given and the rule is learned (p. 4). This is especially the case in Spanish because it is a highly inflectional language, with a fairly regular suffix system for headwords.

According to Nation (2001), languages have a relatively small group of words, or lemmas, that are very frequent. These frequent words are particularly important because they make up very large percentages of written and spoken texts. The general number that has been set for what is considered to be high-frequency is the 2,000 most frequent lemmas. Nation and Hwang (1995) write that the first 1,000 of which covers 77% of the continuous word-forms in American English and 5% more for the second set of one thousand (p. 35). A learner of English, or any language for that matter, would thus greatly benefit from learning such highly frequent words. For the same reasons, a second language learner would suffer greatly in terms of his or her communicative competence if there was a lack of knowledge of these highly frequent words that are going to be encountered in his or her daily life outside of the classroom.

2.3.4 Vocabulary Size

The next logical step is to combine the two ideas of vocabulary size and frequency into a discussion of the ideal vocabulary size of a language learner. Leading to this discussion is the information on lemma frequency as well as the ideas of Nation and Waring (1997), which include how an ESL learner’s vocabulary level should take into consideration the vocabulary use of his or her native-speaking interlocutors. It is obvious that for a learner of English, knowing at least the majority of the first 1,000 most frequent words, while possibly insufficient for communicative competence, is the crucial necessities for a person wanting to become competent in comprehension and production skills (see section 2.3.4).

Carter (1998) goes into further detail about second language vocabulary learning, describing the rate of vocabulary growth generally accepted for second language learners. He describes how learners should learn about 1,000 words a year, while having a two to three thousand word fallback if they want to match the vocabulary growth of an adolescent in his or her native language (p. 236). There are, however, no explicit, agreed upon standards in regards to these numbers. For example, Renouf (1984) studied nine major communicative beginning level EFL textbooks that ranged in total number of word forms from 1,156 to 3,963. This shows that for textbook authors and publishers, there are extremely different opinions about how many words a beginning level student should be exposed to in his or her first course. On one hand, a student might be exposed to a much smaller vocabulary but the quality and use of repetition in that exposure might lead to better long-term retention than a textbook which presents a larger, less repetitious vocabulary. As discussed in section 2.2.1, different approaches may have different beliefs on the ideal size of input for a learner. Textbooks from different approaches would thus reflect such different believes. Exposure or input, however, is critical to the learning process. This is not to generalize that presenting more words is always better. Quality of word choice relies on a number of other factors, including types of words presented, methods of presentation, number of times presented, and integration of the material in the classroom. However, while a learner realistically does not retain in long-term memory all of the input he or she receives, information not presented cannot be learned, even if only to be retained short-term. For example, a learner using the textbook with 1,156 words might learn every one of the words he or she is expected to learn from that particular book; however, if textbooks were the only source of input, a student would not have the opportunity to learn as many vocabulary items from a textbook that presents 1,156 words as a student using the 3,963 word textbook would.

Furthermore, Carter (1988) states that it is generally claimed that if a learner knows the first 2,000 words (at least in English), he or she will have about 80% lexical coverage in a real language environment (p. 236). According to Nation (2001), however, for a learner to comprehend a text well, they need to have about 98% understanding of the words given (p. 114). Hirsh and Nation (1992) found that a vocabulary size of 5,000 was needed to allow for such an understanding, resulting in 98.5% coverage of known words (p. 695). A vocabulary size of 3,000 has been shown to be needed to have a coverage of about 95%, which percentage of known items in a text Liu Na and Nation (1985) determined to be needed to begin to efficiently use context to guess the meanings of unknown words (p. 38).

2.3.5 Word Lists

Using frequency data from corpora, researchers and material developers are able to create lists of important, or highly frequent lemmas. Nation (2001) describes how corpora can also monitor which words are frequent in what types of settings or ranges. To determine these specialized vocabularies, researchers use specialized corpora that consist of instances of the target genre in which the target language is used, giving learners a more specialized vocabulary depending on the purposes for which they want or need to use the target language. This section will describe the history of and current issues regarding word lists and what Carter and McCarthy (1988) refer to as the “vocabulary control movement” (p. 1).

In the 1930s, Ogden (1930, as cited in Carter & McCarthy, 1988) proposed a method of teaching English called Basic English. This method was based on the idea that a learner should know at least the bare essential linguistic (syntactic and vocabulary) knowledge needed to communicate his or her ideas. As Carter (1998) describes, the originators and proponents of this method felt that the learners should not be burdened too much by having to learn extensively large amounts of vocabulary, so instead, learners were taught 850 highly-frequent word forms and only the basic productive rules to minimize learning troubles (pp. 23-28). While this method paved the way for other word-list based pedagogical methods, it was lacking in usable application. For example, the 850 words were not based on data from corpora, so intuition must have played an important part in the list’s development. Also, by limiting one’s learning to 850 words, there could be significant problems for a learner desiring communicative competence (see section 2.3.4). This would be especially problematic in a second language setting in which the learner needed to interact with native speakers who probably would not be familiar with the system of Basic English or know how to “simplify” their own speech significantly for adequate communicative exchanges.

The next major development in the “vocabulary control movement” was Michael West’s *A General Service List* (GSL), published in 1953, containing 2,000 word families. Compared to *Basic English* West’s GSL has had much more durability in the area of language teaching, and has had continued use through the twentieth century (Carter & McCarthy, 1988). The selection of the words on the GSL was based on their frequency as found in a corpus of written English of 2 to 5 million words, as it was continuously modified. Another belief of GSL proponents, as Carter (1998) explains, is that the learner should be told the frequency of the word he or she is learning as well as the relative importance of various meanings a word form might have. As described earlier, Nation and Hwang (1995) showed that these first 2,000 lemmas in English represent about 82% coverage of the running words found in the corpus used (p. 35). One of the problems with the GSL, however, is that it is based on relatively old data. The original corpus and list, over 70 years old, would not represent potentially frequent words in current use of English that refer to concepts that did not exist or were not frequent at the time. Examples might include words that refer to modern innovations like *computer*, which, according to Leech and Wilson's website (n.d.) is the 220th most frequent noun present in the British National Corpus. Another downside to the GSL is that it is based solely on a corpus of written English, possibly neglecting forms frequent in spoken English that do not surface as frequently in the written medium.

In the 1980s, a much more ambitious project was undertaken under the leadership of John Sinclair by the University of Birmingham and what is now the publisher HarperCollins. This group formed the Collins Birmingham University International Language Database (COBUILD) project. This project's goals are to better understand the details of how English is naturally produced and how that can be applied to improve the instruction of English learners. According to the project's website (Collins, n.d.), COBUILD makes use of a corpus of over 524 million words and growing. Carter (1998) summarizes the innovations of the COBUILD dictionaries. He writes that one of the innovations of these dictionaries is the use of contexts based on English that has been spoken and/or written in "real world" situations. This allows the dictionary users to read an example of how an entry naturally occurs in the target language. Carter also describes how materials developed from this project make use of the separation in storage and marking of British English and American English. This separation allows for differences in variations to be accounted for in language research and materials development. Even more innovative, however, is the marking of relative frequency of an entry. The frequency information allows both learners and instructors to know the relative importance of a given word. This may be important when determining if a word is worthwhile to learn or teach. For example, a beginning level teacher might prefer that his or her students not focus too much attention on a vocabulary item that is not likely to be encountered again. This would especially be useful when working with authentic texts whose vocabulary is not controlled for appropriateness. Another benefit of these dictionaries compared to more traditional dictionaries is that they offer concordance advice, showing what forms frequently occur with a given entry. Such concordance information is an integral part of lexical-based approaches and methods that emphasize language "chunks" (see section 2.2.1). Finally, the COBUILD dictionaries and materials also emphasize frequent discourse markers and seemingly content-less words that are frequent in spoken English, but because of their relative absence in written English had been largely neglected by other such dictionaries.

While there are not similar, established dictionaries and word lists in Spanish, they could be created using the same methods. Because accurate frequency dictionaries rely on corpora, and only recently have there been adequate or appropriate Spanish corpora, older Spanish frequency dictionaries have had definite limitations. According to Davies and Face (2006), these dictionaries had all been quite old (most over forty years old) and based on very small corpora (less than three million words). Another problem with these older corpora and frequency lists in Spanish is that they were not always lemmatized. Before the online premier of Davies' *Corpus de Español* [Spanish Corpus] in 2002, not only had there not been any readily accessible corpus close to its size and depth (100 million words from both Spanish and Latin American sources), but also few other corpora had been lemmatized, making grouping of inflected forms difficult. Davies' frequency dictionary, *A Frequency Dictionary of Spanish: Core Vocabulary for Learners* (2006), like the *Corpus de Español*, is based on several small corpora from varying countries (both Spain and Latin America) and sources (spoken, transcripts, literature, and texts) from the last century. With a total of about twenty million running words, this combination of corpora gives a relatively balanced representation of Spanish as a whole, across dialects and genres. From these corpora, the most frequent, or useful, words were derived. Other, newer Spanish dictionaries based on corpora, such as dictionaries based on *Corpus de referencia del Español actual (CREA)* [Reference corpus of modern Spanish] (n.d.) and the Lara's (1996) *Diccionario del español usual en México* [Dictionary of general Mexican Spanish], have used corpora, but frequency is not explicitly addressed by these dictionaries as it is with those created by the COBUILD project. In other words, there is no distinction between highly frequent, moderately frequent, and only partially frequent entries.

2.4 Materials Development and Analysis

This final section of the literature review discusses the pedagogical implications of the previous sections. In particular, it addresses pertinent vocabulary inclusion in syllabuses and materials, such as textbooks, used in language learning environments. Ellis (2001) summarizes his research by describing how a student's input in an instructional setting can be distorted compared to native use of the target language and that this distortion may lead to unnatural patterning and thus frustration. This might especially be more prevalent in second language settings because a student may have much more input and social interaction in the target language outside the classroom. Also, because of the ideal to teach what naturally occurs in a target language or language variation, Gavioli and Aston (2001) describe the need for the planners of language acquisition materials to justify their lexical choices. Such justification, Gavioli and Aston claim, should be given when including a word that is very infrequent or when excluding a word that is very frequent in a large corpora of the target language (p. 239). Corpora and frequency lists, thus offer the needed instruments to determine the appropriateness of that which is or should be included in an second language syllabus.

Sinclair and Renouf (1988) offered what they name *The Lexical Syllabus*. The motivating factor of such a syllabus is that vocabulary should be at least equal if not take precedence over grammar and communicative instruction. This belief in the importance of vocabulary is also reflected in Lewis' (1993) *The Lexical Approach*. In both approaches, the lexicon is seen as holding most of the content or meaning in language, and that through the lexicon, students can learn other aspects of language, such as grammar and communicative skills. Sinclair's COBUILD project (see section 2.3.5) takes particular interest in relative frequency in different target ranges or genres, relying on the very large and sophisticated corpus to make judgments of word choice.

For a concrete example of materials analysis and as a precedent for the study proposed here, Davies and Face (2006) explored the appropriateness of the vocabulary words in SFL textbooks. These researchers investigated six textbooks, three first-year and three second-year Spanish textbooks published and used in the United States of America. They made use of Davies' (2006) frequency dictionary, which is based on corpora from various Spanish-speaking countries and genres and from both written and spoken modalities (see section 2.3.5). Davies and Face compared this list of the 5,000 most frequent lemmas in Spanish to the vocabulary words taught in the textbooks. Instead of studying vocabulary are defined by Sinclair (1991) as all the words presented in a text, Davies and Face (2006) only focused on the active vocabulary. This *active vocabulary*, according to the researchers, represent the vocabulary that the textbook authors generally expect the target language students to learn. This is compared to *passive vocabulary*, which is represented by the words that are present only in context. The authors may not expect students to learn or retain the meanings of these contextual entries in their long-term memories. Davies and Face (2006) collected the active vocabulary by extracting all of the words that were presented out of context from glossaries and word banks. The researchers limited their scope to only active vocabulary to make stronger conclusions about the appropriateness of what students are expected to learn in terms of frequency. By including passive vocabulary into one's research, as the current study does, one cannot make judgments of appropriateness because such a methodology does not allow one to know which presented forms are expected to be learned.

Davies and Face (2006) found that amongst these widely used textbooks, frequent lemmas were significantly neglected. The researchers also found that in terms of percentages, if any one of these textbook presented 2,000 vocabulary items (there was a range of 523 to 3,217 total active vocabulary items), only 10% to 50% of those items would be part of the all-important 2,000 most frequent lemmas in Spanish. This means that at best, of the six textbooks studied, only half of the items most important to speaking and understanding the language would be presented.

These results have particular importance in Spanish language education. As described earlier, much of the research in corpus linguistics and its relation to vocabulary and pedagogy has studied English, thus any further investigation related to other languages is needed. Also, it was important to determine whether there is much of a difference between what the authors of these textbooks studied felt to be important and the actual frequent forms of Spanish in real use. Such a difference shows that not only do more Spanish textbooks need to be examined, but also that the writing of such materials should take into account frequency from the beginning. This is especially the case in a second language environment, where the learner is surrounded by real uses of his or her target language. Thus, there would be a great benefit to studying SSL textbooks as well as textbooks that have been written by native speakers and published in the country in which they are to be used, as there may be instances of variation-specific instruction or the reliance on intuitive beliefs based on anecdotal evidence. Finally, such textbook analyses will offer language professionals, such as teachers, the knowledge of what types and specific examples of vocabulary are neglected by a particular material being used. This would allow them to supplement their instruction, giving their students a stronger base knowledge of the most useful aspects of the target language.