Vocabulary Coverage in Textbooks for Learners of Spanish

Jack A. Hardy

Universidad de las Américas, Puebla

Abstract

This paper describes the investigation and description of the vocabulary in two beginning level, Spanish as a second language textbooks that were both published in Mexico. This investigation makes use of a relatively new method of textbook analysis, which involves the measuring of real-world frequency of the vocabulary presented in a text. A frequency list of the most frequent 5,000 words in an extensive corpus of modern written and spoken Spanish was used to describe the words chosen by the textbooks' authors. The vocabulary in this paper is described in terms of overall coverage of frequent and non-frequent entries. In more specific terms, however, this study also investigates the under- and over-represented entries. These are words that are highly frequent according to the corpus and words that are presented in the textbook but are not in the frequency list, respectively. Because both of these textbooks were written with the second language learner in mind, this paper also describes how Mexican-specific vocabulary is treated by both textbooks.

Table of Contents

Section

1. Introduction

Words are an essential part of language, thus vocabulary is an integral part of second language learning. Language learning and teaching methods are generally based on theories or beliefs about language. According to Pinker's *Words and rules: The ingredients of language* (1999), language is basically constructed of memorized forms and grammar (words and rules). These memorized forms of sound or sign, words, are arbitrarily matched to meaning. In recent theories of second language *acquisition* (to be used interchangeably with *learning* in this thesis), words or lexical entries have gained prominence in the implementation of communicative and lexical approaches. As long as learners' second language lexicons are given importance, second language acquisition (SLA) materials should teach that which is most common in a "real world" setting of the target language, giving the learner the best general base of the language as possible. This should especially be the case in a second language environment in which the learner is living amongst speakers of the target language and needs to be familiar with frequent lexical entries in order to successfully communicate.

This thesis is based on an exploratory investigation of the "real world" frequency of the word-forms presented in two Spanish second language textbooks. It is a preliminary study, making use of a relatively new method of textbook analysis. Thus, this study does not attempt to make any judgments of textbook quality, but instead describes these materials with the hope that future studies will improve on the methodology used. There is also the possibility that such studies will be used as ways to help analyze the overall quality of textbooks, helping create better materials for language students.

1.1 Purpose and Overview

The purpose of this study is to make use of a new method of textbook analysis to study vocabulary coverage. Two beginning level Spanish as a Second Language (SSL) textbooks are examined. These Mexican-published works are Duhne, Emilsson, Montoya, and del Río's seventh edition of *Pido la palabra: 1er nivel* [I call for the floor: First level] (1998) and Canuto, Cortés, Escobar, Gutiérrez, and Montemayor's second edition of *¡Estoy listo!: Nivel 1* [I am ready: Level 1] (2003). The central objectives of this study are to describe and analyze the vocabulary from these textbooks. Sinclair (1991) describes *vocabulary* as the overall number of different words in a text (p. 29). The 60vocabulary from these two textbooks was described in terms of their frequency levels in authentic Spanish. In other words, this research studies all of the words in two textbooks that are directed towards students as these words relate to their frequency in spoken and written Spanish.

To determine frequency and coverage of frequency ranges, the vocabulary from these textbooks was extracted and compared to Davies' (2006) lexical frequency list of Spanish based on a large corpus. This corpus, the Corpus del Español (n.d.) represents both speech and writing from Spain, Mexico, Central America, South America, and the Caribbean (Texts section). Using the frequency dictionary developed by Davies, these data were then described and analyzed in various manners, including frequency range, relative coverage compared to total number of vocabulary entries, and syntactic category. Data from a recently published study by Davies and Face (2006) gave methodological groundwork as well as data from American-published Spanish as a Foreign Language (SFL) textbooks. This frequency list was used to determine different quantities and percentages of word-forms presented in the textbooks as well as what frequent Spanish word-forms were not presented by the textbooks. For example, two of the overall questions

raised were how many and what kinds of words in the frequency list were not present in the textbooks. Conversely, the infrequent entries that the textbooks did present were also explored for their relative percentage amongst all the number of words presented in the textbooks as well as by their syntactic categories. Finally, the data obtained from these two SSL textbooks were also compared to the first-year, Spanish as a Foreign Language (SFL) textbooks studied by Davies and Face (2006).

1.2 Rationale for this Study

The elements of vocabulary frequency and coverage explored (see section 1.1) represent the basic questions of this study. However, it is also important to understand why this study was carried out. One motivating factor for this study was that most of the research that has been done on target vocabulary in terms of frequency has been performed on English. What little research has been done using large corpus-based frequency lists has also primarily investigated English as a Second Language (ESL) and English as a Foreign Language (EFL). Furthermore, there is a gap in the research literature on how the frequency of vocabulary is considered in textbooks and other materials used to help language teaching when taught as a second versus a foreign language. By investigating Spanish instructional materials, this study offers preliminary data into an emerging field in second language acquisition and corpus linguistics, where data from multiple languages can be triangulated for more universal theories. Spanish is a particularly appropriate language to be studied not only because it is a different language than English but also because of its number of speakers and learners. According to Gordon (2005), Spanish is a widely spoken language across the globe: spoken by around 322 million native and 60 million second language speakers (Spanish section, para. 1).

Another motivating factor for this research is that, according to Davies and Face (2006), no corpus of Spanish larger than a million words had been made publicly available until 2001 (p. 2). Therefore, it would have been difficult for research to be conducted on the appropriateness of the vocabulary a learner is expected to know in terms of frequency of use in an authentic environment of the target language. Thus, preliminary and descriptive studies like this one may lead to further research. Such investigations will use methodological precedence from exploratory studies to make use of corpora and frequency lists not only to help determine the appropriateness of the vocabulary in language textbooks, but also help to make advances in textbook analysis and creation. For example, by studying textbooks individually in the proposed manner, a language program director could know before he or she implements materials what issues regarding vocabulary coverage might arise if a particular textbook were to be adopted in his or her program. By knowing this information beforehand, a director could then inform the teachers using the textbook of potential problems. For example, the program could use data from a frequency list derived from corpora of the target variation, whether it be standard or dialectal, to determine potential missing pieces to its materials. Administrators could thus provide a list of word-forms in the top 500 frequent words in the target variation that are not covered in the textbook. This would allow the teachers to know how, specifically, they could supplement the given materials. Appendix A offers such example lists developed for the two textbooks investigated in this study as they relate to Davies' (2006) broad frequency list that represents several regional variations of both written and spoken forms of Spanish. However, because of the finite amount of time a teacher can spend with his or her students, one may not expect all of these entries to be taught. Such a list may be more useful as a guide for a teacher to see where he or she could supplement existing lesson plans,

especially if some of the entries fit into themes or situations that has already been designed in the syllabus.

1.3 Significance of this Study

This study offers an exploratory glimpse into one aspect of the vocabulary in two second-language textbooks. Because most of the research has been done on English education, studies like this could help the process of better textbook design, particularly in languages that have been relatively understudied in terms of pedagogical implementations of corpus linguistics studies. In the case of Spanish, as such a largely spoken and taught language around the world (Gordon, 2005, para. 1), it is particularly important that research is executed specifically on it and not simply relating findings based from research on English. However, such methodologies and theories about English language acquisition can be investigated relative to other languages in attempts to make methodological and theoretical improvements in the overall field of vocabulary learning and second language acquisition.

Textbooks offer an ideal source of preliminary corpus linguistics research because of their permanence. Textbooks give learners a written source of the target language. This modality not only allows learners to go back and revisit the language that they are exposed to but also provides a major source of input for dozens to tens-of-thousands of learners. With the potential to be one of the sole permanent and easily retrievable sources of a learner's target language input, choosing what is to be presented to so many learners should be taken seriously.

The researcher realizes that vocabulary is only one aspect of textbook design, which is only one aspect of a language course, which is yet another piece of the overall curriculum

design. However, as Richards and Rogers (2001) point out, more emphasis is being placed on the mental lexicon in theoretical linguistics with even Chomsky (2000) giving more importance to the lexicon and semantics in grammar theories (pp. 169-173). Similarly, applied linguists, including Sinclair and Renouf (1988), Lewis (1993) and Nation (2001), have also brought more attention to the importance of mental lexicon's role in a language learner's overall communicative ability in the target language.

Finally, with advances in computational abilities, corpus linguistics has allowed linguists and lexicographers to catalog millions of real utterances in any given language (Richards and Rogers, 2001). Now that the technology and materials from these advances (such as frequency lists) are available, new research can be carried out to first examine and then to evaluate language-teaching materials. The methodology used by Davies and Face (2006) and repeated in this study could become a particularly useful research methodology to help understand and then maintain an authentic and appropriate balance between what is produced in the "real world" of the target language with what a learner can be expected to understand and acquire. Because the corpus used is representative of so many regional variations of Spanish, it is not necessarily a mirror for any one context for a learner. Instead, "real world" here refers more to texts produced by fluent speakers of the target language that are used to better understand natural language production. According to website for the Corpus del Español (n.d.), the corpus used to create the frequency list, not only accounted for regional differences it also controls for genre. The modern section of the corpus, which was used for the frequency list, equally represents "literature, oral texts, and newspapers/encyclopedias" (Texts section).

1.4 Theoretical Framework

In general, the theoretical framework of this study is quantitative. According to Ellis (1999), studies like this one, using language corpora, are generally observational and quantitative in nature (pp. 31-33). As such, the study was designed to better understand or describe already existing materials, and the researcher's possible influence on the data and results is not being explicitly examined. Gay and Airasian (2002) write that this type of quantitative research, studying current status or pre-existing data, is called survey or descriptive research (p. 10). The difference between qualitative studies that try to describe current statuses like existing materials and quantitative studies with the same goals, these authors write, is that data collected in quantitative survey research are categorized in terms of numbers instead of more open-ended answers like narratives. The numbers in quantitative studies like this one are usually fixed and unable to be manipulated in the phase of data collection. In the case of the present study, all word forms from the textbooks were extracted and then assigned frequency numbers based on a fixed frequency list. There was no noticeable way for the researcher to manipulate these assignments. Even compared to other quantitative, experimental or semi-experimental studies, the current study does not make use of data trimming.

While described as quantitative because it relates its data in terms of fixed numbers, descriptive corpus linguistics neither represents an extreme version of positivism nor the most quantitative of the research methods (experimental). Instead, according to Larsen-Freeman and Long (1991), the method of focused description is directly in the middle of a continuum of qualitative to quantitative research methods (p. 15). It is the researcher's belief that the method of data collection used in the current study is not as easily influenced by the researcher himself as in either of the extreme ends of qualitative or quantitative research. For example, in corpus linguistics studies, such as in the current one, there is no

discrimination in the word-forms that are included. However, as to be further discussed in the chapter on the methodology used in this study, especially in the sections on limitations and delimitations (see section 3.4), it is made clear that the data derived from these entries are not representations of a single truth. This research is observational and descriptive, and in a post-positivist context, its goal is not to determine the definite quality of the materials being studied.

1.5 Definitions of key terms

This study uses a few key terms that need to be defined or operationalized. The following is a list of basic definitions and descriptions of some important key terms that are found through much of this thesis.

- *word-form*: Sinclair (1991) defines a word-form as “an unbroken succession of letters” (p. 28).
- *vocabulary*: Sinclair (1991) defines vocabulary as all of the different word-forms presented in a text (p. 29).
- *active vocabulary*: Davies and Face (2006) define active vocabulary as “the vocabulary that students are expected to learn and be able to use, and is generally the vocabulary included in chapter vocabulary lists” (p. 4).
- *passive vocabulary*: Davies and Face (2006) define passive vocabulary in terms of the texts of materials used to teach a second or foreign language. They describe such vocabulary as “words that appear in the text, often in reading passages, which may be glossed so that students can better understand the content that they are

reading, but these words are not meant to be learned and used by students at this point" (p. 4).

- *lemma*: According to Nation (2001), a lemma as the base form of a word and its inflected variations (p. 7). There are various ways to operationalize a lemma. This study operationalized lemmas in the same way as Davies and Face (2006). They describe a lemma as consisting of a headword and its inflected forms (p. 5). If two word-forms are spelled the same way but of different syntactic categories, they are considered to be two different lemmas.e-language
- *"real world"* and *authentic*: These terms will be used to describe written or spoken texts that have been produced by fluent speakers in a natural environment, not necessarily intended to be used in a second or foreign language learning context.
- *Communicative competence*: According to Canale and Swaine (1980), "communicative competence is composed minimally of grammatical competence, sociolinguistic competence, and communication strategies" (p. 27). All of these areas are seen as important for a second language learner to successfully interact with speakers of the target language.

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.

3. Methodology

3.1 Overview

As discussed in Chapter 1, the overall design of this project involved the comparison of the vocabulary in Spanish as a Second Language (SSL) textbooks to a frequency list developed from corpora of usage by native Spanish speakers. The two books studied were *Pido la palabra: Primer nivel* (1998) and *¡Estoy listo!: Nivel 1* (2003). As a conceptual replication of Davies and Face (2006), this project used similar methods in an attempt to answer research questions regarding the vocabulary choices made by the designers of Spanish-language textbooks. The questions investigated were specified as:

- How well represented are frequent lemmas, as determined by a frequency dictionary, in these Spanish language textbooks?
- What kinds and to what extent are vocabulary items under-represented and over-represented in these textbooks?
- Are there any noticeable differences or similarities between the vocabulary coverage by these second language textbooks and the foreign language textbooks studied by Davis and Face (2006)?

While this study is principally investigative in nature, there are some particular hypotheses regarding the research questions posited. These hypotheses can be described in terms of possible outcomes. For example, before completing the study, the researcher hypothesized that there would not be that many differences between the results in the current study and those of Davies and Face (2006). Especially because Davies and Face found such wide variety in the vocabulary coverage amongst SFL textbooks alone, there was not expected to be a large difference in coverage or word-types (in under- and over-

representation) between these two SSL textbooks and the textbooks that Davies and Face studied. Based on the Davies and Face (2006) findings, it was also predicted that there would be decent but not complete coverage of highly frequent words, even though native speakers design the books for an audience in a second language environment. One might expect Level 1 books to include the most frequent content words. However, when only intuition or traditional themes are used to determine vocabulary choice, some frequent lemmas might be neglected. In English, for example, as found by Biber and Reppen (2002) the first 12 most common verbs make up 45% of the use of all lexical verbs (p. 205). Even though such verbs are obviously very important in communication, according to these researchers' findings, textbooks for beginners disregarded many of these words (pp. 205-206). Thus, another hypothesized outcome was that these first year textbooks in Spanish also would lack some of these highly frequent and useful words. However, the current study analyzes the vocabulary of the textbooks as a whole and not only the active vocabulary (see section 3.3.1), such frequent function words might be in the final list of extracted vocabulary even though those items are only presented passively.

Investigating these questions contributes to both an emerging methodology for analysis of textbooks in the hopes of improving pedagogical materials. Such studies are important because of the lack of research on languages other than English as well as a call for an improvement of available instruments. For example, continued research in this area could lead to more interest, funding, and innovations in the way that corpora and frequency lists are created, managed, and used. This study and others like it are also important for pedagogical reasons. Currently, the method of analyzing vocabulary with accurate frequency lists is neither commonplace in the analysis of nor in the creation of materials for Spanish language teaching. As discussed in section 2.3.2 on intuition, even native speakers

are not always good judges of frequency. As described earlier in the description of the study by Biber and Reppen (2002), five of the twelve most frequent lexical verbs were found to be entirely neglected in a series of first-level ESL books gives even further justification for studies like this one.

The following sections go into detail on the background of the current study and how it was executed. In section 3.2, the discussion of materials includes the textbooks being investigated and the frequency list and corpora used. This section is followed by a discussion on procedures, in which vocabulary extraction, lemmatization, and frequency assignment are described (section 3.3). Following these general methodological descriptions of the study, section 3.4 discusses assumptions, limitations, delimitations and other methodological questions.

3.2 Materials

3.2.1 Textbooks

Two first-year Spanish as a Second Language (SSL) textbooks, *Pido la palabra: Primer nivel* (1998) and *¡Estoy listo!: Nivel 1* (2003), were examined. The Universidad Nacional Autónoma de México (UNAM) in Mexico City published both, and both were created through the UNAM's *Centro de Enseñanza para Extranjeros* [Center for the Teaching of Foreigners] (CEPE). Because Davies and Face (2006) researched Spanish instruction books published in the United States for the use of foreign language learners, a replication of their study could benefit our understanding of word choice and coverage by studying books written by different authors for different purposes and targeted towards a different audience. The two books being studied for this project were chosen because they

were written by native Spanish speakers, were published in Mexico, and are widely used in Mexico to teach SSL.

Another difference between these two books and those studied by Davies and Face is the intended audience. The textbooks analyzed in this study were written to target second language learners who are studying in a Spanish-speaking country (Mexico). According to the introduction of *¡Estoy listo!* (2003), both it and *Pido la palabra* (1998) were designed with the Examen de Posesión de la Lengua Española [Test of Spanish Language Proficiency] (EPLE) in mind (p. 12). This test, according to the CEPE is designed to measure the proficiencies of foreigners interested in studying Spanish as second language at the collegiate level in Latin America. Thus, even if these materials were to be used in a foreign language environment, their designs would still reflect second language goals. As a second language audience, the students would generally come from different cultural and linguistic backgrounds. Also, the target language would be based on a target culture. In this case, Mexican Spanish and Mexican culture. In a foreign language environment, there might be much more homogeny amongst the students with instruction could integrate and compare the target language and culture to those of the students.

Finally, these particular textbooks were singled out because they are widely used. According to the preface of *Pido la palabra* (1998), these two books are used in more than 130 institutions around the world, including Mexico (p. presentación [preface]). *Pido la palabra*, in particular, is easily accessible even in small bookstores in central Mexico. *¡Estoy listo!* (2003), on the other hand, may be less common as it was not available at the same small bookstores in provincial Mexico. However, the researcher quickly found available copies at the UNAM bookstore and a large commercial bookstore in Mexico City. This is understandable as *¡Estoy listo!*, while used as an SSL textbook (p. 10), is at the

same time oxymoronically described by the authors as a Spanish textbook for a non-immersion environment (p. 11).

3.2.1.1 *Pido la palabra*

Pido la palabra: *Primer nivel* (1998) is the first in a series of five textbooks designed to teach non-native Spanish speaking foreigners how to speak Spanish in the Latin American environment. The first edition of this book was notably published in 1988 when work on lexical importance and emphasis by researchers like Sinclair and Nation was only in its infancy. In fact, this was the same year that Sinclair and Renouf (1988) published their pioneering work in the "vocabulary control movement," *The Lexical Syllabus*. *Pido la palabra* was first written in a time when a strict version of the Communicative Approach to language teaching dominated the field.

According to *Pido la palabra*'s (1998) introduction, the main objective of this textbook is to present linguistic and communicative aspects of Spanish for the situations second language students are likely to encounter in their daily lives in a Latin American environment. The textbook is divided into 13 units, each centered on such common situations. Each unit begins with a synopsis of the learning objectives for that unit, described in terms of thematic/social content, communicative objectives, and linguistic content (see Appendix B for excerpts from the Table of Contents, showing a typical unit and all 13 units' topics and listed vocabulary themes). Throughout the textbook, the designers of *Pido la palabra* also labeled the exercises based on the tasks required to complete them. Listening comprehension, oral expression, reading comprehension, written expression, and critical thinking or reasoning are the tasks described. In terms of the design

of the textbook, *Pido la palabra* contains 282 instructional pages, is written in black and blue inks, and has graphics (both in color and in black and white) on nearly every page.

The writers describe the ideal use of the book to be in an intensive, 60-hour, six-week course (p. X) where students are immersed entirely in Spanish in a communicative naturalistic environment, supplementing the learning that takes place while living in a second language environment. This textbook is regularly used in both private and public universities throughout Mexico to teach Spanish to speakers of other languages who are living in Mexico. However, according to language teachers familiar with using the *Pido la palabra* series, these textbooks are also regularly used as the college-level textbooks for a three to four hours per week, semester-long classes.

In keeping with the Communicative Approach (p. IX), the authors refer to communicative competence, authentic materials, strategies, inductive learning, and interaction in their introduction (see Appendix C for the authors' list of methodological bases). However, their only reference to vocabulary is in how the book is structured. Vocabulary is included as part of the linguistic content needed for the topics covered by each unit. These communicative priorities of the authors may have influenced the frequency or appropriateness of the vocabulary, as well as the manner that vocabulary is presented. For example, in this textbook new vocabulary is rarely treated as a separate entity from other grammatical lessons. There are very few word banks or vocabulary lists, and there is no glossary. With 282 pages, there is, however, a very large amount of vocabulary, although not necessarily active vocabulary.

This large amount of vocabulary in readings and the lack of explicit instruction coincides with communicative as well as natural approaches which emphasize the importance of sufficient input in the target language and inductive learning (see section

2.2.1). As described in the introduction of *Pido la palabra* (1998), the authors do not expect the target learners to understand all of the input they receive, but be able to understand what is important and to grasp main ideas (p. X). This textbook may thus not be ideal for a vocabulary study designed to measure quality of coverage. The researcher is unable to determine which words are expected to be understood, learned, or skimmed over. This difference from the more modern, lexical textbooks in the Davies and Face (2006) study, led to a change in which vocabulary items in this study would be extracted (see section 3.3.1). Instead of only extracting active vocabulary, the current study examines all vocabulary presented by the textbooks. Because these differences in the textbook design decrease the ability to make judgments of appropriateness, the current research is more of an exploratory description of the vocabulary already chosen by the authors than in judging the quality of the textbooks.

3.2.1.2 *¡Estoy listo!*

Although *¡Estoy listo!: Nivel 1* (2003) may not be designed for a learner entirely immersed in Spanish (p. 15), it is widely used across Mexico to teach Spanish to speakers of other languages while living in a Spanish-speaking country (p. 11). In this way, although not necessarily an exclusively second-language textbook, it is regularly used as such. Furthermore, the directions in this book are all written in Spanish, and the only foreign-language aspect that the writers implemented in its design was to add glossaries with English and French translations of vocabulary words. In terms of the authors' beliefs of how languages should be taught and learned, they write that there are three main aspects: communicative, grammatical, and lexical knowledge (pp. 15-18). This is an eclectic mix of various approaches described above, in which language is seen as being composed of

multiple aspects, and not one over others. These basic beliefs are reflected in the design of the textbook. For example, the authors believe that lexical content is of the same level of importance as communicative and grammatical content, so it is given a more important role in this textbook compared to *Pido la palabra* (1998). Because of this belief and the more recent publication of *¡Estoy listo!* the authors may have also been more aware of choosing appropriate target vocabulary and decontextualizing these target lexical items. To help students gain these three types of linguistic knowledge, the authors write in their introduction that oral and writing production are given the same importance as listening and reading comprehension (p. 17). The chapters use various exercises to help develop these four skills. *¡Estoy listo!* consists of five, situationally-based units. Each of these units has specific communicative, grammatical, and lexical goals (see Appendix D for excerpts from the Table of Contents, showing a typical unit and all 5 units' topics and lexical objectives).

Compared to *Pido la palabra* (1998), *¡Estoy listo!* (2003) has more of a workbook style. *Pido la palabra*, consistent with the Communicative Approach, emphasizes inductive learning and provides readers with a lot of input.*¡Estoy listo!*, on the other hand, consists of mostly pictures, word banks, short dialogues, and fill-in-the-blank exercises. The length of *¡Estoy listo!* (280 pages) is comparable to *Pido la palabra*, but the font is significantly larger in the prior, and there are very few large blocks of continuously running text. Interestingly, while *Pido la palabra*'s authors were clear to mention that their book's readings were almost entirely authentic, *¡Estoy listo!* appears to be almost the opposite, consisting of short dialogues and readings, apparently targeted specifically towards low-proficiency learners. The preface of the textbook describes how vocabulary is presented in and that grammar is taught through simple, understandable context (p. 16). Such a

structural difference, of preferring constructed readings and activities to authentic contexts, might be due to the fact that it was also designed for foreign language instruction.

3.2.2 Frequency List

Besides the textbooks, another important instrument in this study was a frequency list with which the two textbooks' vocabulary coverage could be compared. The frequency list used was Davies' (2006) *A Frequency Dictionary of Spanish: Core Vocabulary for Learners*. As described earlier (sections 1.5, 2.3.5), this is a list derived from a representative combination of corpora from a variety of countries, modalities, and genres.

Because this project focused on SSL as taught in a Mexican environment, there could be some conflict comparing the frequency of vocabulary taught in a Mexican SSL textbook with a frequency list based on worldwide Spanish. However, the materials available determined the manner in which the study was executed (see section 3.4.2 on the limitations of this study). For example, while it would have been ideal to compare the textbooks' vocabulary exclusively to lexical frequency in Mexican Spanish, the materials available for this variety do not match the combination of size and depth of Davies' corpus and frequency dictionary. Not only is this corpus large, but also unlike the even larger CREA corpus, the entries of Davies' corpus are lemmatized and categorized for collocations and syntactic properties. While researchers have long used Spanish corpora for lexicographic studies, most of that research has emphasized overall description and has not necessarily focused on frequency. The goals and academic projects for using the CREA of the Real Academia Española (n.d.), for example, are prescriptive in nature. An example of this is that according to the Real Academia Española's website, the mission of these projects is to "avoid changes in the Spanish language and the constant evolution so that the

unity between speakers of Spanish is maintained." When the CREA debuted online for public use, one would have hoped it could have been used by outside researchers for frequency studies. However, according to Davies and Face (2006), this 120 million-word corpus was neither annotated for part of speech nor was it lemmatized (p. 2)

Meanwhile, in terms of exclusively Mexican Spanish, Lara (1990) has worked on the lexicography in more descriptive manner. The *Corpus del Español Méxicano Contemporáneo* [Corpus of Contemporary Mexican Spanish] (CEMC), of which he is the director, is of a decent size and country specific; however, at less than two million running words, it is not much larger than those Spanish corpora used fifty years ago (Davies and Face, 2006, p. 3). Also the organization of published materials of frequency derived from this corpus is not conducive to textbook analysis as the published works derived from this corpus have generally not included specific frequency assignments. Similar to the CREA, the goals of this corpus are more conducive with lexicography than with corpus linguistics. That is to say, that frequency is not explicit in published studies and materials.

There is also the number of words in a frequency list to consider when analyzing textbooks. For example, a frequency list of the most frequent one hundred words in a language may not be very useful in the analysis of a textbook that contains 3,000 word-forms. Available Mexican Spanish lists from the CEMC only include around the first two thousand words. While a first year Spanish student may benefit from only learning the first 2,000 most frequent words, such a list would only allow a researcher to investigate the coverage of highly frequent (#1-2000) and not moderately frequent (#2001-5000) entries. Davies and Face (2006) found that in first and second year SFL textbooks, there are a significant amount of vocabulary words in the frequency ranges between 2001 and 5000 (1205 lemmas across all six textbooks, or 40.2% of the total 3,000 items that could

potentially be represented from that range). While Davies and Face do not offer data on individual textbooks on this question, the textbooks in the current study also present a significant amount of vocabulary in the 2001-5000 range, as shown later in Chapter 4, Results and Analysis. Of the total number of lemmas presented in *Pido la palabra* (1998), 26.33% were found in this range. In *¡Estoy listo!* (2003), 24.97% of the total lemmas presented come from the same range.

As a supplementary tool, however, such Mexican Spanish frequency lists or dictionary entries could help understand any noticeable dialectal differences unique to Mexican Spanish that might be present in the textbooks. An example of such would be the textbooks' omission of the verb *coger* in Mexican Spanish books. It is a relatively frequent verb in some dialects of Spanish with a frequency number of 1896 in Davies' (2006) list, meaning, "to hold, take, catch." According to the *Pocket Oxford Spanish Dictionary* (2003), however, its use in Mexican Spanish is limited to a vulgar meaning. Thus, one might not expect such a word to be taught in a first-year textbook that is designed for learners of Spanish in Mexico. On the other hand, there could be entries in textbooks that might be frequent in Mexico but infrequent in other Spanish dialects. Variation specificity was taken into consideration in the labeling of entries in order to help to realize any outlying data. This process is further discussed in upcoming section 3.3.3, Lemmatization.

Another way that the materials available influenced or limited the methodology of this study is that such frequency lists as those of Davies (2006) and Lara (1990) generally only take into account orthographic words. That is to say that even these recently-created frequency lists do not yet allow for easy comparison of multi-word lexical entries with an easily accessible measure of collocation. An example of how this limits understanding of the lexicon is that some lexical items such as idiomatic expressions or verb phrases like

echar a perder [to rot] have different meanings than the sums of their parts. Thus, measuring each orthographic word might not reflect the frequency of certain frequent word combinations. With the influence of applied linguists like Sinclair and Renouf (1988), Willis (1990), and Lewis (1993), language teaching programs have begun to focus on communicative and lexical approaches in which a common methodology is for the student to often learn entire phrases or "chunks" of language. This aspect of word "chunks" or collocations in language teaching and learning, however, would be difficult to measure using orthographic-word-based frequency lists. Further work is needed in the development and publication of frequency lists of Spanish. Lists that take into account frequent word collocations, or "chunks," for example, would allow for more accurate descriptions and investigations of lexical entries, and not just orthographic words. Further corpus linguistic studies investigating topics of frequency in textbooks could also, in turn, improve the implementation of such second language acquisition approaches with more lexical emphases in languages other than English.

3.2.3 Corpora and Dictionaries

With the recent advent and availability of Spanish corpora comparable to the large, established corpora in English, the corpus linguistics findings and theories of West, Sinclair, and others can now start to be applied towards Spanish. The principal interest of this particular study was not in using corpora, but rather in making use of a frequency list obtained from corpora. However, it is important to understand the corpus used to create Davies' (2006) Spanish frequency dictionary as well as the dictionaries referenced in order to better understand the data collected.

One such corpus is Davies' *Corpus del Español* (2002). With over 100 million running words in Spanish, this corpus can be divided into sections of historical eras. The section of this corpus which this research used is that of Modern Spanish (the last century), of which there are over 20 million words. This more modern section was that used to create Davies' (2006) frequency dictionary. This corpus was also used in the current research as a supplement to Davies' frequency list. In order to generalize the coverage of an infrequent word, Davies and Face (2006) entered the lemma into the corpus search engine to determine its number of total occurrences in the corpus. Although not thoroughly investigated in this study, a similar process was used to show just how infrequent some of the words presented by the textbooks are (see section 4.3).

Although not used directly in this study, another application of this corpus' website could be to investigate collocations. While one cannot determine the frequency of a group of words like the phrasal lexical entry *por supuesto* [of course], its relative frequency can be investigated using this corpus. For example, one can do a search for *por* and solicit the environments in which it occurs. Through this, one can determine how common *supuesto* is in relation to the first word. Another option that this site gives is to search an entire phrase. Again, this will not give the researcher a frequency number, *per se*, but it will show how many instances that phrase was encountered in the given number of total words searched, from which a percentage of frequency could be derived. In the case of *por supuesto*, both orthographic words are listed in the frequency dictionary. While *por* occurs in a wide variety of environments, *supuesto* relies much more heavily on the preposition. The corpus, for example, shows that in 66 of 100 random contexts of *supuesto* in 1900's Spanish, the word was preceded by *por*. Davies (2006) addresses common collocations by listing the

phrases next to the entry when the lemma occurs in that phrase in a significantly sizable amount of the total number of that lemma's occurrences (p. 9).

Besides Davies' (2006) frequency dictionary, three other dictionaries of Spanish were also consulted for meaning and dialect appropriateness. The first, the *Pocket Oxford Spanish Dictionary* (2003) was used as a general tool to obtain short definitions for words not present in Davies' (2006) dictionary. It was also used to determine if entries not present in the frequency dictionary were exclusive to Spanish spoken in Mexico and/or Latin America. Entries that were specific to the region were labeled as such, making this dictionary an easy reference for regional variation. This dictionary was also used because at 90,000 entries, it is a relatively extensive pocket dictionary. This would be big enough to explain most infrequent words that might not be present in a phrase book, but small enough to be for a second language learner to use for reasons of portability to and from class, and in everyday life since the students are living in a second language environment. In other words, it is this researcher's belief that a student in his or her first Spanish class should be able to find the vocabulary presented in that course in this type of dictionary without having to resort to the consultation of a large desk dictionary. Supplementally, a much larger dictionary, *Simon & Schuster's International Spanish Dictionary: Second Edition* (1998), was consulted for words used in the textbooks but not present in either of the previously mentioned dictionaries.

The third dictionary consulted was Lara's (1993) *Diccionario fundamental del español de México* [Fundamental dictionary of Mexican Spanish]. This was used to determine how many of the dialectally Mexican and Latin American Spanish entries were important or useful enough to be placed in a list of the top 2,500 most essential Mexican Spanish lemmas as determined by El Colegio de México's *Corpus del Español Méxicano*

Contemporáneo [Corpus of Contemporary Mexican Spanish] (CEMC). This dictionary, according to Lara's (1996) introduction in the *Diccionario del español usual en México* [Dictionary of general Mexican Spanish], contains the lemmas needed to basically understand general or scholarly texts like the textbooks examined in this study. Some of the common *mexicanismos* [words important and/or specific to Mexican Spanish] presented by these textbooks and also present in this dictionary include *cheque* [(bank) check], *chile* [chile, peppar], *frijol* [bean], *jitomate* [tomato], and *platicar* [to talk, chat].

However, the same dictionary also included seemingly obscure or rare entries like *chahuiztle* [mold, plague] and *chapopote* [tar] and did not represent relatively more everyday Mexican Spanish words like *ahorita* [right now], *enojado* [angry], or *mesero* [waiter]. Such a discrepancy might exist because this dictionary and other frequency lists are not always based solely on overall frequency. The list makers can also take into account the amount of different types of texts in which an entry surfaces. If an entry surfaces hundreds of times in only one source, it might not be as important in the overall frequency as a word that surfaces a few times in every source. This process of weighting was used both by Lara (1993) and Davies (2006), in attempts to create frequency lists more reflective of speech and writing as a whole.

Finally, one important source for information on Mexican Spanish will not be used. Lara's (1996) *Diccionario del español usual en México* [Dictionary of general Mexican Spanish] is much larger than the fundamental version. It is so extensive; however, that it contains nearly all of the Mexican variation lemmas found in the SSL textbooks in this study. The use of such a general Mexican Spanish dictionary would shed little light onto a lemma's frequency as it contains a large number (around 14,000) of lemmas without reference to their comparative frequency.

3.3 Procedure

The procedure for this investigation followed the same basic steps as those performed by Davies and Face (2006), but it only investigated first-year textbooks. From these textbooks, vocabulary items in the form of orthographic words were extracted, lemmatized, and entered into a spreadsheet, where they were labeled in terms of frequency number. From this point, the words were placed into bands of frequency to better understand the vocabulary coverage of both textbooks, to compare the information between the textbooks, and to find any possible similarities or differences between the first-year textbooks examined by Davies and Face and those in this study.

3.3.1 Vocabulary Extraction

The first step of the procedure was the extraction of the vocabulary from the two textbooks being studied. In the Davies and Face (2006) study, all of the textbooks in question decontextualized their vocabulary in what the researchers labeled active vocabulary. Such vocabulary is called active because they are the words that the textbook writers generally expect the students to learn and be able to produce. The design of all six of the textbooks investigated happened to include easily accessible lists of these words in the forms of word banks and glossaries.

In this study of these SSL textbooks, however, only one of the textbooks (*¡Estoy listo!* (2003)) presents vocabulary in such lists. The active vocabulary in *Pido la palabra* (1998), on the other hand, was not clearly available. *Pido la palabra* lacks any form of glossary, and the word banks utilized are few and far between. Also, some of the frequently used words in the various contexts given are not presented out-of-context.

The focus of this study would have ideally been of decontextualized entries, as was the case in the Davies and Face (2006) study. Examples of decontextualization in these textbooks include word lists, words matched with pictures, expected production of the word, and other activities using the word outside or in multiple contexts. However, although, the authors of *Pido la palabra* mention in their table of contents the semantic groups of vocabulary expected to be learned for each chapter, this does not show exactly which words the students are expected to learn in terms of vocabulary (especially when it came to function words). Also, although there are target vocabulary themes for each chapter, the vocabulary associated with those themes is not always presented out of context in those chapters. Perhaps because of the highly communicative, almost naturalistic approach of this particular textbook, the methodology of word extraction was significantly changed.

Because of the difficulty in determining what was meant as target vocabulary in one of the textbooks, the methodology for vocabulary extraction was changed from that used by Davies and Face (2006). Instead of only using decontextualized vocabulary, all of the orthographic words in the textbooks after the introductions, which explain the textbooks to the language program directors and teachers, were entered. This slightly changes assumptions that can be made on expectations of learning. For example, one of the more frequent verbs, *dar* [to give] (number 39 in frequency) is only given in context in two situations in *¡Estoy listo!* (2003). This lemma was thus entered into the data, but a learner might not, necessarily, learn it. Interestingly, however, the total of entries extracted from the two books being studied was not far off from the textbooks in the study being replicated. In the Davies and Face (2006) study, the average first year textbook contained

2,317 lemmas that were either in book final glossaries or presented out-of-context. In the present study, 2,175 is the average number of total lemmas presented in the two textbooks.

The first step of extraction was the copying of individual, orthographic words as they appeared in the textbooks. Again, because of the communicative nature of *Pido la palabra* (1998), the strictly active vocabulary was too difficult to be determined. Thus, the nature of these results of this study is not directly comparable to those of the Davies and Face (2006) study (see section 3.4.2 on these limitations). The orthographic words were entered into a spreadsheet, and the syntactic category with a simple definition was placed to the side of each entry to help the researcher remember what the word meant in the context in which it was presented. If a word was not present in Davies' (2006) frequency dictionary or in the *Pocket Oxford Spanish Dictionary* (2003), the page number of where it could be found was placed instead of a definition. This is shown in Table 1.

Table 1.

Excerpt from orthographic word entry spreadsheet

Lemma	Syntactic Category	Definition
diamante	n	diamond
charol	n	patent leather
fondo	n	p. 230
combinar	v	to combine

With the page number present, the researcher was able to later confirm the appropriate meaning when looking for a definition in the larger Spanish dictionary by *Simon & Schuster* (1998). The words were also entered in order of appearance, so when the researcher needed to refer to how a word was presented in the textbook, even if the

definition was known, he was able to later find the page(s) where that word had been presented. Thus, the second step to the extraction of orthographic words was the deletion of repeated entries. Identical orthographic words were deleted if they shared the same part of speech. If there were two words spelled the same, but of different syntactic categories, both were kept. Because each orthographic word in the textbooks was entered, there were several multiple entries. This particular study is not investigating the number of instances an entry is presented in a text or the quality of that presentation. Instead, any presentation of a word was used, only showing the existence of the vocabulary item in the material.

3.3.2 Lemmatization

Once the vocabulary entries had been extracted, they were categorized and coded to match the forms that are used by the frequency list. This process is called lemmatization. A lemma is a way of describing the basic form of a word (see section 2.3.3). Researchers use these forms to measure vocabulary knowledge, assuming that a learner will also learn the morphological patterns of inflection to create the various other forms of the same syntactic category. Lemmatization allows various forms of a "word" to be studied as a whole instead of each inflection of the lemma counting as a separate entity. Nation (2001) describes how using the lemma as a basis for counting forms in corpora has been used for over sixty years, making lemmatization a standard procedure in research in corpus linguistics (p. 7).

In the same way Davies and Face (2006) processed their extracted vocabulary, there are two main types of lemmatization that were utilized: one for verbs, and another for nouns and adjectives. Basically, lemmas are determined by the types of affixes that the individual orthographic words contain. If a group of words all have the same base but differ

only in inflectional affixes, they are of the same lemma. These differences do not reflect a change in the part of speech amongst the different forms of the lemma.

The treatment for the lemmatization of verbs is straightforward: the infinitive form was entered into the lemmatized vocabulary list. Also, the adjectival forms of verbs were treated as adjectives. In Spanish, these are usually the words in which the infinitive of the verb is altered with a suffix of *–ido* or *–ado*. For example, *dormir* [to sleep] is given a different entry than *dormido* [asleep] because they belong to different syntactic categories. Thus, they were treated as different lemmas in the lemmatization process.

For Spanish nouns, there is the question of number and gender. In Spanish, most nouns can be singular and plural, with the morpheme /-s/ marking plurality. As a matter of ease, the singular form was used, not only to compare to the singular forms used in the frequency list, but also because the singular is a default, from which the learners would be taught the rule to pluralize. However, there were discrepancies about how the textbooks actually present the vocabulary. For example, one of the vocabulary words given was *recámaras* [bedrooms], and its singular equivalent was never given. In order to compare such a word to the frequency list, its lemma (the singular form) was used. Another potential concern regarding pluralization was whether or not a singular word and its equivalent ending with *-s* were actually forms of the same lemma. This was based on the entries' meanings. The example given by Davies and Face (2006) was that of *botones* [buttons, bellhop, bellhops]. In that particular textbook, only the latter meanings were given, thus the singular *botón* [button] was not included in the lemmatized vocabulary list because of its difference in meaning. Similarly, for homonyms, words that are spelled and pronounced the same yet have different meanings and/or syntactic categories (i.e. *ayuda* [help, aid] vs. *ayuda* [to help, 3rd person, present, indicative]), the appropriate lemma was decided

depending on which syntactic category the book uses. It is of note, however, that the frequency list does not distinguish between different meanings of a homonym of the same syntactic category. For example, there is one entry in Davies' (2006) frequency dictionary for the noun *palma* even though it has two very different conceptual meanings: [palm of a hand] and [palm tree].

In terms of gender, when both masculine and feminine forms of a noun exist and have the same meaning (except gender assignment), the masculine was chosen to represent the lemma as a whole. An example of such an occurrence is *abogado* [male lawyer] and *abogada* [female lawyer]. As seen in plurality, differences in gender in this sense do not change the syntactic category of the entry. Because of this, such pairs were generally entered as a single lemma. Adjectives that can take both masculine and feminine endings depending on the gender of the word to which they refer were treated the same way. This research is not making any claims into which gender would be marked and which is unmarked. Instead, this is merely a way in which to combine the two forms in order to evaluate the frequency of these various forms as a single lemma. However, in the frequency list there were some lemmas that differed only in terms of gender, such as *hijo* [sg., son; pl., children] and *hija* [daughter], yet these pairs were given two separate frequency assignments in Davies (2006) Spanish frequency dictionary, which possibly causes some inconsistencies. Other examples of such feminine nouns that are frequent enough to warrant recognition in the top 5,000 most frequent Spanish words include familial words such as *prima* [female cousin], *tía* [aunt], and *niña* [girl]. Davies does not explain why this difference is noted in familiar lemmas and not others. Because there was no way to determine how frequent the two entries would be combined, they were also given separate

lemma assignments in this study. Thus, when making lemma assignments, the frequency dictionary had to be consulted for any feminine nouns presented in the textbooks.

3.3.3 Frequency Assignment

While vocabulary words were being lemmatized, they were arranged alphabetically in spreadsheets: one spreadsheet for each textbook's vocabulary. They were arranged in alphabetical order because Davies (2006) arranged half of his dictionary by frequency and the other by alphabetical order. Alphabetical order allowed for easy data entry of frequency numbers and verification of syntactic categories and definitions. The syntactic category information proved useful to better understand which types of words the authors presented because of the possibility of homonyms of different syntactic categories, allowing for the appropriate frequency assignment.

Once a lemmatized vocabulary list was created for both of the textbooks being studied, the lemmas were assigned a number based on their positions in the frequency list of Davies' (2006) Spanish frequency dictionary. Going through the alphabetical list, words were assigned a frequency assignment with one being the most frequent and 5,000 being the least frequent. Because the frequency dictionary gives the first 5,000 most frequent lemmas in Spanish, any less frequent words presented in the textbooks were not assigned a frequency number. Through access to this frequency list, the entries were assigned simple numbers and not a coverage percentage. When an entry was not present in the frequency list, the word was looked up in the *Pocket Oxford Spanish Dictionary* (2003) for a definition and possible variation assignment (see section 3.2.3). After being looked up, the definition and syntactic category were then written in the columns next to the lemma's entry, but no frequency assignment was given. In the dictionary used, each entry is

evaluated on its dialectal appropriateness. This was particularly useful for better understanding a second language textbook from Mexico. When a word was used either in Latin America (AmL) or Mexico (Mex) exclusively, the researcher was able to better understand why such a lemma would be present in a Mexican textbook but not present in a frequency list of Spanish across dialects. This information was then transferred to the entry by writing "(Mex)" before the definition. Finally, if a lemma was not present in this smaller dictionary, the more extensive dictionary was consulted. These entries were then placed in bold, allowing the researcher to know which lemmas needed to be further investigated.[1] Below, Table 2 shows examples of the alphabetical lists made for the two spreadsheets.

[1] Although not directly part of the study, there were significant (although not large) amounts of these bold entries, being present in neither the frequency dictionary nor the pocket dictionary. In *Pido la palabra* (1998), 84 of the 2924 presented lemmas (2.87%) had to be looked up in the much larger dictionary. *¡Estoy listo!* (2003) has a similar coverage of such entries: 31 out of 1438 (2.15%). These numbers are similar to those of Mexican and Latin American specific vocabulary (see section 4.4)

Table 2.

Excerpt from lemmatization spreadsheet

Lemma	Syntactic Category	Definition	Frequency
a	prep	to, at	5
abogado	n	lawyer	1680
abreviatura	n	abbreviation	
abrigo	n	overcoat, shelter	2996
.	.	.	
libar	**v**	**to taste, drink, sip**	
.	.	.	
rentar	v	(Mex) to rent	

A copy of the two master list spreadsheets was made from which to arrange the data. As in the Davies and Face (2006) study, repetitions and any proper nouns and numbers that were not present in the first 5,000 most frequent word list were deleted, giving a final count of the general vocabulary used in the textbooks.

Once those proper nouns and numbers were eliminated, and once all of the other lemmas had been assigned a frequency number or had been determined not frequent enough to receive one, the data were then arranged in different spreadsheets. One spreadsheet for each textbook was used to order the entries based on their assigned frequency number. Next, on separate spreadsheets the items were categorized by both frequency and syntactic category. This means that all of the lemmas of the same syntactic category could be combined, and in those subsets, the lemmas could be ordered by frequency. This ability to rearrange and group the data based on frequency, syntactic

category, and a combination of the two allowed the researcher to analyze different aspects of the data to be discussed in Chapter 4, Results.

3.4 Other Methodological Topics

The methodological precedents for this particular study are discussed in the literature review in the description of the Davies and Face (2006) study, which is being replicated. The basic method, as described in detail in the previous sections, was based directly on the methods put forth by Davies and Face. The procedures of textbook selection, lemmatization, and frequency assignment were directly adapted to the study of Mexican SSL materials. However, this study differs in the method of vocabulary extraction from that of the study being replicated. Instead of studying active vocabulary, this study investigates the presentation of all the orthographic words in the textbooks with the exceptions of proper nouns and infrequent numbers Other examples of such feminine nouns that are frequent enough to warrant recognition in the top 5,000 most frequent Spanish words include familial words such as *prima* [female cousin], *tía* [aunt], and *niña* [girl]. The further sections on assumptions, limitations, and further questions discuss the relative scope of and potential drawbacks to the methodology of this study.

3.4.1 Assumptions

The methodological assumptions here refer to the learners, the textbooks, and the study being replicated. For example, the researcher assumes that nearly all of the vocabulary in the textbooks being used will be new to the learners. This is assumed because the textbooks being studied are designed for non-native Spanish-speaking learners who have no significant background knowledge of Spanish. This assumption, however, may not

be accurate, as this study does not investigate students who use these materials. Neither the actual learning of the vocabulary nor the depth in which those items are learned, were taken into account in the current study. With the exploratory and descriptive methodology used on only the materials, there is no way to clearly determine the rate of which learners actually do learn these particular words by studying the texts alone. This is important because it means that the researcher cannot judge the quality of the textbook; he can only describe the raw data. Future, mixed-methods studies might be able to shed more light onto the overall picture of vocabulary learning as it relates to textbooks, methods, activities, attitudes, etc. Another assumption of the researcher is that the frequency list created by Davies (2006) is an accurate reflection of the corpus he used and that the corpora he used to extract those lemmas are an accurate representation of modern Spanish across country and dialectal boundaries.

3.4.2 Limitations of this Study

The limitations of a study are the uncontrollable or unexpected variables that the researcher encounters that may affect his or her study. One of the major limitations to this particular study is in the area of the materials. For example, textbooks studied by Davies and Face (2006) were all explicit about what vocabulary items were expected to be learned because of their presentation in word banks, glossaries, or other out-of-context situations. This active vocabulary allowed the researchers to make claims about the appropriateness or quality of the word choice. In the current study, however, one of the textbooks being studied is not, particularly, designed for explicit vocabulary learning. *Pido la palabra* (1998) does not make much of a distinction between active and passive vocabulary. For example, there are very few word banks and no glossary. Also, several function words are

never explicitly presented out of context, but by the number of times they are presented, target learners would probably be expected to learn them.

Besides the textbooks, the instrument used to determine vocabulary entries' frequencies was also a limitation. This study focuses on textbooks designed for second language learners, learning in an environment where the target language and variation of that language are being spoken. Ideally, such a study would use a frequency list derived from a corpus of that particular variation. According to Ham Chande (1979), the Colegio de México's *Corpus del Español Méxicano Contemporáneo* [Corpus of Contemporary Mexican Spanish] (CEMC) has the relatively small size of just under two million tokens (individual, orthographic words) across genres and ranges in Mexican Spanish. However, the resources published from it have been more lexicographic, creating dictionaries, than frequency related. Examples of published works include the *Diccionario fundamental del español de México* [Fundamental dictionary of Mexican Spanish] (1993) and the *Diccionario del español usual en México* [Dictionary of general Mexican Spanish] (1996) (see section 3.2.3 for the description and limitations of these materials).

Davies' (2006) frequency dictionary, while not exclusively Mexican Spanish, covers a wide variety of variations of Spanish, including that of Mexico. This dictionary was also derived from a much larger corpus (about 20 million tokens) than the CEMC. Also, as Moreno de Alba (2005) describes, the fundamental, or frequent, words across variations of Spanish do not differ very much. Moreno de Alba specifically refers to the 1,451 most frequent lemmas in the CEMC, representing 75% of all Spanish utterances. Nearly all of these lemmas, he claims, correspond to general Spanish across variations and that very few would be specific to Mexican Spanish. Finally, as a frequency list, Davies' frequency dictionary is more user-friendly for investigation purposes. There are three sets

of lists, ordering the 5,000 most frequent words in different orders: by frequency, alphabetically, and by syntactic category and frequency. This allowed easy access to item information and more than twice the amount of entries than available sources based on the CEMC. To curb the possible effects that using a non-variation specific frequency list would have on the results, entries that were primarily of Mexican or Latin American variations were tagged to later be compared to data from the Mexican Spanish derived CEMC (see sections 3.2.3, 3.3.3). Finally, another limitation to Davies' (2006) frequency dictionary is that it does not distinguish between different meanings of a homonym of the same syntactic category. For example there is one frequency entry for *pila* that includes its various meanings of baptismal font, battery, and heap. These are all different concepts, and it is not probable that a low-level student would have such a deep knowledge of individual vocabulary entries.

3.4.3 Delimitations of this Study

The delimitations of a study reference that which the researcher has determined to be the scope of the study. Delimitations allow the researcher to focus on a particular area of interest. In the post-positivist era, it is important to recognize that even quantitative research with raw data and numbers does not represent an entire truth. By recognizing that there are other aspects to second language acquisition and vocabulary learning, the researcher can qualify his results, allowing his or her readers to better understand their place in the field. For example, this study is an exploratory and descriptive investigation of vocabulary frequency, and it is not particularly interested in the manner of presentation of those vocabulary items. Thus, one of the largest delimitations of this study is that the study is only interested in any presentation of vocabulary. Not only does it not take into account

the way that such items are presented, but it also does not study how often an item is presented. According to Nation (2001), both of these aspects are important factors to the successful acquisition of vocabulary. Nor does this study investigate the order in which vocabulary entries are presented or the methodology implemented in the classroom. Because of these reasons, the current study does not make claims about the quality of either of the textbooks, as there are more aspects to language and vocabulary acquisition than those studied here.

3.4.4 Further Discussion of Methodological Questions

Some methodological questions remain. To begin with, the research design is one of investigating the presentation, based on frequency, of vocabulary in two textbooks. This methodology has some obvious limitations. For example, analysis of these textbooks does not take into account the learning of a student from his or her teacher(s), peers, or environment. Another aspect of vocabulary learning that will not be addressed directly by this study is that of frequency within the textbook. For example, one vocabulary word might appear once in a short lesson, never to be used again in the textbook, and possibly not by the learner. On the other hand, there might be a vocabulary entry that is taught early in the textbook and reused in various receptive and productive contexts. The more times an entry is encountered, the more likely the learner is to be able to understand and produce, showing a deeper knowledge of the word and its uses. Thus, instead of being a study of first-year SSL learners' vocabulary levels, this project is merely focusing on one aspect of the vocabulary learning process, the coverage of vocabulary items used in textbooks. Another reason the methodology of this study lends itself more to studying materials

instead of learners is that one cannot be sure by only analyzing a textbook if learners really do learn the words that are targeted by the books as important.

This methodology also lacks the ability to measure or describe the use of lexical chunks or phrases that are learned as a whole instead of as separate parts (see section 3.2.2). Especially as these books are both designed in the desire to help foreign speakers learn Spanish in a second language environment, there may be more of these lexical chunks than in a foreign language textbook. Again, as the instruments for investigation improve, it would behoove Spanish language instructors and planners to further investigate such issues. This is especially the case with *Pido la palabra* (1998) that is explicitly based on the Communicative Approach. Not only are vocabulary words in this text rarely overtly pointed out, but also the majority of the exercises are based on conversations. Such conversations may, in accordance with the Communicative Approach, be used to gain communicative competence and an understanding of the main ideas in a conversation. This method may be ideal for these goals, but there would not be a way to determine which individual words the students are learning merely by examining the textbook. The instruments themselves also affect how such chunks can be studied. As described above, the frequency list being utilized is based on orthographic lemmas. That is to say that multiple word lemmas are not taken into account. While one cannot directly compare such phrases being taught, as described earlier with the example of *echar a perder*, corpora can be consulted to find the commonality of collocations between certain words compared to their overall use.

Finally, the instrument itself was not the ideal one for this study. Because these textbooks are designed to teach second language students Mexican Spanish in Mexico, the ideal frequency list to be used would be based solely on Mexican Spanish. Davies and Face

(2006) used an ideal frequency list, based on several variations and genres of Spanish because they were studying foreign language learners, who would probably want to learn the broadest uses of Spanish instead of any particular variation. Due to reasons of size in Mexican Spanish corpora (less than 3 million words) and frequency lists (2,000 words), however, the Davies' (2006) frequency dictionary was chosen. While there were not many dialectal differences amongst the 5,000 most frequent words, the outliers that surface in the data when using Davies' dictionary were secondarily triangulated with data from the Mexican Spanish data in the corpora. These particular words were found by the (Mex) added to the entries that were said to be used more or less exclusively in Mexico or Latin America (see section 3.3.3).

4. Results and Analysis

4.1 Overall Coverage

The first set of data to be presented and analyzed is the overall coverage of the vocabulary in terms of frequency ranges in Davies' (2006) frequency dictionary. Table 3 gives information of the number of lemmas that each textbook presented for each of the ten, 500 lemma ranges as determined by the frequency dictionary. The coverage, the percentage out of a total of 500 for each range, of each textbook is provided to the right of the amount of lemmas in that respective range.

Table 3.

Overall coverage of top 5,000 lemmas by range and textbook

Range	*Pido la palabra*		*¡Estoy listo!*	
	no.	%	no.	%
500	442	88.4	313	62.6
1000	340	68	181	36.2
1500	289	57.8	157	31.4
2000	217	43.4	109	21.8
2500	202	40.4	95	19
3000	155	31	78	15.6
3500	143	28.6	73	14.6
4000	118	23.6	37	7.4
4500	78	15.6	43	8.6
5000	74	14.8	33	6.6
TOTAL	2058	41.2	1119	22.4

The final line represents the total number of the 5,000 most frequent lemmas represented by the textbooks and the respective overall coverage of those 5,000 most frequent lemmas.

Table 3 shows that the first range, which represents the 500 most frequent lemmas, is well covered by both textbooks. *Pido la palabra* (1998), for example lacks only 58 of the words in this first frequency range. For a beginning level course, this high level of coverage makes sense. Without much, if any background in Spanish, the target learners need to have the basics of the language to successfully communicate in a second language environment. Also, as Lara (1993) claims, the first 1,451 words in Mexican Spanish represent a coverage of 75% of all cultural linguistic utterances (p. 11). Through the third range where this number lies, *Pido la palabra* presents more than half of the 1,500 lemmas. 1,071 of these lemmas are presented with a coverage of 71.4%. *¡Estoy listo!* (2003), on the other hand, presents 651, or 43.4%, of the same top three ranges. In terms of total numbers, *Pido la palabra* presents almost twice as many of the top 5,000 lemmas than *¡Estoy listo!*.

While *¡Estoy listo!* (2003) presents only slightly more than half of the total lemmas that *Pido la palabra* does, it shows similar priority towards the top ranges in relation to less frequent ranges. All three of the top three ranges of 500 cover a larger percentage of words than the overall coverage of 22.4% of the most frequent 5,000 lemmas. This is similar to the results of Davies and Face (2006), which found that the further down the scale of ranges, the less coverage there was. Besides this general trend, these data cannot be directly compared to the Davies and Face study because their article does not distinguish between textbooks at this level of coverage. Instead the researchers compiled all of the words from each level (first and second year), as if a learner were to use all three of the textbooks of one level at one time.

The current study also compared the total number of words that were presented in each textbook with how many of those words were in the list of the top 5,000 lemmas. Table 4 shows the total number of lemmas presented in the textbooks, the number of lemmas presented that are in the frequency dictionary, the number of lemmas presented that are not in the frequency dictionary, and the percentage of in-dictionary lemmas relative to the total number of lemmas presented.

Table 4.

Coverage by textbook: percentage of words in frequency dictionary

	Total no. lemmas	no. + dictionary	no. - dictionary	% + dictionary
Pido la palabra	2924	2058	866	70.38
¡Estoy listo!	1438	1119	319	77.82

The data here show that the majority of the vocabulary presented in both textbooks is frequent enough to be in the frequency dictionary. Although *¡Estoy listo!* (2003) does not present as many total lemmas as *Pido la palabra* (1998), the vocabulary that it does present is more likely to be encountered by the target learners in the "real world." This could be due to the fact that the linguistic input for learners presented in *¡Estoy listo!* was written in a possibly more controlled way than the authentic readers of *Pido la palabra*. Authentic texts, like those in *Pido la palabra*, might be more likely to present more infrequent words than texts written with a beginning level Spanish learner in mind.

The Davies and Face (2006) study also showed these percentages of words in first year SFL language textbooks. The total numbers of all active vocabulary lemmas in the first year textbooks were 2,218, 1,616, and 3,217. The percentages of these words that were also in the frequency dictionary were 85%, 81%, and 78%, respectively. All three of these

figures are higher percentages than *¡Estoy listo!* and *Pido la palabra*. However, these comparisons across the SSL and SFL textbooks are not completely valid because Davies and Face only extracted active vocabulary, and the current study extracted all presented vocabulary. It might be the case, for instance, that had the passive vocabulary been included in the Davies and Face study, more infrequent lemmas would be included, lowering this percentage. On the other hand, had only active vocabulary been extracted in the two SSL textbooks studied, the amount of total lemmas from *¡Estoy listo!* would have been even less. Also, although *¡Estoy listo!* has a relatively high coverage rate, it only presents a total of 1,438 lemmas. This is significant because it shows that the textbook may not be providing enough input for the student and may need to be lexically supplemented by other materials. According to Renouf's (1984) study of EFL textbooks (as cited in Sinclair & Renouf, 1988), a textbook with 1,438 total lemmas would be at or near the bottom of the list of amount of lemmas presented. Such a comparison should not be seen as totally valid, however, because in this type of textbook analysis one can neither know how much is done with the vocabulary presented, how often it is used in the classroom, nor how well it is learned.

Another way to examine a textbook's vocabulary coverage was proposed by Davies and Face (2006). This measurement of coverage examines coverage in a different way. These researchers give the example of a quantitatively ideal textbook:

> Suppose that a textbook has N number of words, e.g. 1,300 words. In the "best of all worlds" scenario, these 1,300 words would correspond to words #1-1,300 in the frequency dictionary. In other words, it would be as though the textbook vocabulary corresponded exactly to the listing in the dictionary. (p. 8)

The total numbers of vocabulary entries for both textbooks in this study are shown above in Table 4. *Pido la palabra* (1998) presents a total of 2,924 lemmas. Of those, 1,619 correspond to the words 1-2,924 in the frequency dictionary. This means that 55.37% of the 2,924 lemmas in *Pido la palabra* relate to words 1-2,924 in the frequency dictionary. *¡Estoy listo!* presents a total of 1,438 lemmas, and 634 of them correspond to the words 1-1,438 in the frequency dictionary. The coverage in this case is 44.09%. These numbers are significant because they mean that 44.63% and 55.81% of the lemmas in these respective textbooks do not relate to the first 2,924 and 1,438 words, respectively, in the frequency dictionary. This means that for the number of lemmas that each textbook presents, they both have relatively low coverage of the most frequent lemmas up to those respective numbers, which shows that infrequent entries may be taught at the expense of more frequent entries. It should be noted, however, that this is an artificial construct of the 'ideal' vocabulary in a textbook. This construct only addresses frequency and neglects aspects of semantic fields or themes. In terms of affective factors, teaching only frequent words, which include most function words, might be boring for both a teacher and his or her students.

Davies' (2005) claims that the first 1,000 most frequent lemmas in Spanish constitute up to 80% of written and 88% of spoken Spanish. This number is slightly different than Lara's (1993) assertion that the first 1,451 words represent 75% of Spanish, possibly due to differences in lemmatization, available corpus data, and frequency calculation. However, either way, it is clear from both of these sources that the first 2,000 most frequent lemmas are critical to producing and understanding Spanish. If nearly half or more of those critical entries are not presented to a first-year student, there could be critical problems dealing with everyday communication.

The previous tables, however, do not describe what kinds of lemmas are being presented and at what frequency. Similar to Table 4 one can also determine the same raw number coverage based on syntactic category. Table 5 relates the total amounts of lemmas with the amount of top 5,000 lemmas that each textbook presents in respect to syntactic categories.

Table 5.

Coverage by textbook and syntactic category: percentage of words in frequency dictionary

Pido la Palabra

	Total no. lemmas	no. + dictionary	no. - dictionary	% + dictionary	% - dictionary
Nouns	1570	1030	540	65.61	34.39
Verbs	521	435	86	83.49	16.51
Adjectives	624	404	220	64.74	35.26
Adverbs	101	90	11	89.11	10.89

¡Estoy Listo!

	Total no. lemmas	no. + dictionary	no. - dictionary	% + dictionary	% - dictionary
Nouns	797	584	213	73.27	26.73
Verbs	214	199	15	92.99	7.01
Adjectives	289	208	81	71.97	28.03
Adverbs	56	47	9	83.93	16.07

This table shows that in both textbooks, the verbs and adverbs that are presented are much more likely to be frequent than the nouns and adjectives that are presented. In the case of *Pido la palabra* (1998), nouns and adjectives are both more than twice as likely to be too

infrequent to be present in Davies' (2006) 5,000 lemma frequency list than verbs and adverbs. It is shown here that the number of nouns presented for both textbooks is more than the total of all the other syntactic categories combined. It is also of note the percentages of adverbs and verbs that *¡Estoy listo!* (2003) presents that are in the frequency dictionary. This shows that, although only a small number of adverbs and verbs were found in the textbooks, the ones that were presented were very likely to be frequent. Nouns and adjectives may have more likely to be concrete and fit better into a themed chapter than verbs or adverbs. This difference may possibly cause textbook authors to use more infrequent, theme-specific nouns and adjectives to fit existing conceptual ideas for a lesson.

The data from Table 5 may be potentially misleading, however, because they refer to a raw numbers and not relative coverages. In terms of raw numbers, *Pido la palabra* (1998) presents 940 more nouns than adverbs from list of the top 5,000 lemmas. This difference may not be relevant in the discussion of syntactic category coverage, because there are many more nouns in the frequency list than verbs. Table 6 represents the overall coverage of content-word lemmas (nouns, verbs, adjectives, and adverbs) across the ten different frequency ranges, relative to the total number of those respective items in the frequency list as categorized by syntactic category. These percentages compare the number of in-dictionary entries presented with the number of that respective syntactic category in the top 5,000 lemmas. For example, of 1,030 of the 2,511 nouns (85.98 %) in the top 5,000 lemmas were presented by *Pido la palabra* (1998).

Table 6.

Vocabulary coverage in percentage by frequency range and syntactic category

Pido la palabra					*¡Estoy listo!*				
Range	N	V	Adj	Adv	Range	N	V	Adj	Adv
500	85.98	90.44	84.88	93.33	500	57.93	63.97	62.79	57.78
1000	70.89	68.75	62.82	42.31	1000	43.04	29.86	28.21	19.23
1500	60.64	49.58	59.43	60.00	1500	34.14	19.33	37.74	25.00
2000	46.25	35.59	47.75	13.33	2000	25.30	15.25	20.72	13.33
2500	42.56	32.56	37.14	46.15	2500	21.45	6.98	17.14	23.08
3000	35.50	20.59	30.09	17.65	3000	19.08	7.84	13.27	11.76
3500	29.15	23.66	28.07	33.33	3500	18.82	5.38	11.40	0.00
4000	25.86	18.69	21.50	30.43	4000	8.37	4.67	7.48	8.70
4500	16.47	17.39	14.58	0.00	4500	12.45	3.26	6.25	0.00
5000	17.88	6.33	13.39	11.11	5000	8.03	1.27	4.72	11.11
TOTAL	41.02	40.43	37.03	43.69	TOTAL	23.26	18.49	19.07	22.82

According to the total coverage, both *Pido la palabra* (1998) and *¡Estoy listo!* (2003) are relatively consistent in representation across syntactic categories. The difference between the percentages of the most-covered and the least-covered categories (adverbs and adjectives, respectively) is only 6.66% and 4.77%, respectively.

4.2 Under-representation

This section discusses the lemmas that were under-represented in *Pido la palabra* (1998) and *¡Estoy listo!* (2003). This particular analysis is needed to better understand what

kinds of words are frequent in native Spanish speech and writing, but are not represented by the first year SSL textbooks. As shown in Table 3, *Pido la palabra* (1998) and *¡Estoy listo!* (2003) cover 442 and 313 of the first range of 500 in the frequency list, respectively. This means that the textbooks do not represent 58 and 187 of the very frequent lemmas (see Appendix A for complete list of these under-represented, highly frequent entries).

To better understand the under-representation of syntactic categories for each textbook, Table 6 can also be read to show the percentages of frequent lemmas that are not covered. For example, *¡Estoy listo!* (2003) only presents 18.49% of the total number of verbs in the most frequent 5,000 lemmas, meaning that 81.51% of the frequent verbs are not represented. Table 7 uses these data to show the total numbers and the percentage of under-represented frequent lemmas.

Table 7.

Under-representation based on syntactic category

Pido la Palabra

	Total no. lemmas in top 5,000	No. of top 5,000 lemmas NOT presented in textbook	Percentage NOT represented
Nouns	2511	1481	58.98 %
Verbs	1076	641	59.57 %
Adjectives	1091	687	62.97 %
Adverbs	206	119	56.31 %

¡Estoy Listo!

	Total no. lemmas in top 5,000	No. of top 5,000 lemmas NOT presented in textbook	Percentage NOT represented
Nouns	2511	1927	76.74 %
Verbs	1076	877	81.51 %
Adjectives	1091	883	80.94 %
Adverbs	206	159	77.18 %

Table 7 shows both how although there are large differences in the number of entries presented in terms of syntactic category and how the coverage of those syntactic categories is relatively consistent for each textbook.[1] In the Davies and Face (2006) study, adverbs were determined to have significantly less coverage than nouns, verbs, and adjectives. Interestingly, however, adverbs in the two textbooks in this study have the best and second-best coverage of the content-word syntactic categories. This may be unexpected not only

[1] Refer to Appendix E for a graphical representation of the segmentation of the top 5,000 lemmas

because of the results from the replicated study in which adverbs were found to be significantly under-represented (p. 9) but also because most adverbs in Spanish, like adjectival forms of verbs, can easily be formed using a simple suffixation rule: adjective + *mente*. According to Davies' (2006) frequency dictionary, of the 206 adverbs in the most frequent 5,000 lemmas, 116 are of this particular composition. If a part of speech had to be more under-represented than the others as active vocabulary, it thus would make sense for it to be adverbs. However, the presentation of adverbs might be common in passive vocabulary because a student may easily understand their meaning even if never presented out of context as long as the adjectival base was already known.

4.3 Over-representation

Another way to describe the lemmas that were presented by these Spanish language textbooks is to show what kinds of words were over-represented. This is an important measurement, as a textbook author should want to be efficient with the vocabulary presented. It would not benefit students to spend time and energy learning infrequent words at the expense of a significant amount of unrepresented frequent words. In this case, over-represented lemmas were operationalized as those not present in Davies' (2006) frequency dictionary. Table 4 shows how many of the total number of lemmas presented in the textbooks were also found in the frequency dictionary. Those lemmas that were in the textbooks but not in the frequency dictionary were considered over-represented. Also, the final column in Table 6 describes the percentages of words, based on their syntactic category, that were presented by the textbooks but are not found in the frequency dictionary. Of all the nouns, for example, that were presented in *Pido la palabra* (1998), 34.39% were not present in the frequency dictionary. In both textbooks, nouns and

adjectives represent the syntactic categories with the highest rate of not being covered in the dictionary.

Another way to better understand the over-represented lemmas in the textbooks is to separate the lemmas that are not in the dictionary and then divide them into their respective syntactic categories. This allows one to see the relative spread of the syntactic categories of the lemmas that were over-represented. It is of note that these data do not compare with any frequency assignment. It is unknown, for example, whether an entry in this section has a frequency assignment of 5,001 or 12,000. Table 8 shows the numbers of over-represented entries in terms of syntactic categories as well as their relative coverage compared to the other syntactic categories. Refer to Appendix F for a graphical representation of this table.

Table 8.

Over-represented lemmas

	Total no. -dictionary	Nouns		Verbs		Adjectives		Adverbs	
		no.	%	no.	%	no.	%	no.	%
Pido la palabra	857	540	63.01	86	10.04	220	25.67	11	1.28
¡Estoy listo!	318	213	66.98	15	4.72	81	25.47	9	2.83

This table shows that nouns are clearly more over-represented than the other syntactic categories. Davies and Face (2006) also found nouns to be much more over-represented than other syntactic categories in their study. One possible reason for nouns and adjectives being more likely to be over-represented is that they hold more obvious, teachable content and fit well into thematic chapters. A physical object, for example, can easily be seen as a picture or object, and its descriptions can be pointed to. This might lead textbook writers to use more infrequent nouns and adjectives because they can easily be taught visually. These

infrequent lemmas might also make for a more interesting lesson, in which students can learn different lemmas that are associated with semantic fields that are familiar to them like parties, food, clothing, and furniture. Verbs and adverbs, on the other hand, may be more difficult to teach because they are not as visually concrete or as easily exemplified physically.

Also similar to the results of Davies and Face (2006), the nouns that were over-represented in these textbooks tend to refer to concrete concepts. These researchers operationalized infrequent entries as not occurring more than 100 times in Davies' (2002) 20-million word, *Corpus del Español*. This definition will be used to describe the extent of infrequency of example entries. Examples of over-represented concrete concepts in the textbooks in this study include *chuparrosa* [hummingbird] (0 occurrences in Davies' (2002) corpus), *hyena* [hyena] (13), *tornasol* [sunflower] (5), and *ventanal* [large window] (79).

Davies and Face (2006) only found four over-represented nouns and two verbs that could be considered abstract. While most of the over-represented lemmas found in *Pido la palabra* (1998) and *¡Estoy listo!* (2003) are similarly concrete, there are also many more abstract concepts represented than in the Davies and Face study. For example, a*dverbio* [adverb] (27 occurrences in Davies' 2002 corpus), *agrado* [charm, affability] (97), *atribución* [attribution] (41), and *astucia* [astuteness] (94) represent four such abstract words from the A's alone. The presence of so many more abstract concepts in this study might be due to the fact that all presented lemmas were extracted from the textbooks instead of only active vocabulary. Similarly over-represented abstract lemmas might also be present but not active in the textbooks studied by Davies and Face.

4.4 Mexican Vocabulary

Entries that were not present in Davies' (2006) frequency dictionary were investigated further in more extensive, bilingual dictionaries (see section 3.3.3). If an entry in these bilingual dictionaries mentioned that a word was particularly used in Mexican or Latin American variations, it was coded with "(Mex)" preceding its definition. This allowed the researcher to determine to what extent these texts were dialect specific as well as to help balance the fact that a multi-dialectal, Spain-dominated corpus was used to create the frequency dictionary. Also, because the target learners for the textbooks in this study are second language learners in Mexico, one might think that much of the vocabulary presented would be Mexico-specific. Table 9 represents the number of these Mexican or Latin American vocabulary entries relative to the total number of lemmas extracted from the textbooks.

Table 9.

Percentage of Mexican/Latin American lemmas

	Total no. of lemmas	no. of (Mex) lemmas	% of total entries
Pido la palabra	2924	97	3.32
¡Estoy listo!	1438	49	3.41

At 3.32% and 3.41% of their total vocabulary being specifically of Mexican or Latin American varieties of Spanish, neither textbook strays too far away from vocabulary that is frequent across all Spanish dialects. This is significant in terms of the question regarding the differences between SFL and SSL textbooks. In the case of these second language textbooks, there is only a small percentage of vocabulary particularly exclusive to Mexico or Latin America.

Because the entries were tagged for general dialect during the frequency assignment process, they could be compared to Lara's (1993) generally frequency-based *Diccionario fundamental del español de México* [Fundamental dictionary of Mexican Spanish]. Table 10 represents how many of the words originally shown to be Mexican or Latin American in the textbooks were represented in Lara's dictionary.

Table 10.

Coverage of (Mex) lemmas in Mexican Spanish dictionary

	Total no. of (Mex)	no. in Mexican Spanish Dictionary	% of (Mex) lemmas
Pido la palabra	97	15	15.46
¡Estoy listo!	49	10	20.41

The data here show that of the total Mexican- and Latin American-specific entries, a relatively small portion of them were frequent in Mexican Spanish. This gives some support to Moreno de Alba's (2005) claim that there would not be many different, variation-specific lemmas in a frequency list of a given variation of Spanish that would not be present in a frequency list based on Spanish as a whole. That is to say, of the Mexican-specific entries that were presented, only a small percentage of them (15.46% and 20.41%) were determined to be frequent in the variation as a whole. However, this dictionary used to determine frequency in Mexican Spanish may not be an appropriate or reliable instrument for such comparisons to Mexican Spanish as a whole (see sections 3.2.3, 3.4.2).

4.5 Summary of results

This section summarizes the results and addresses the initial questions posited above (see section 3.1). The first of these questions asked about how well did these two

SSL textbooks represent frequent lemmas. This question was answered in two ways; the first described how both textbooks represented the first 500 most frequent lemmas relatively well (see section 4.1). However, using another method in which the total number of lemmas that were presented was used as a baseline for the cut-off number in the frequency list, only 55.37% and 44.09% of the N number of lemmas presented are amongst the 2,924 and 1,438 most frequent lemmas in *¡Estoy listo!* (2003) and *Pido la palabra* (1998), respectively. This shows significant room for possible coverage improvement in these two Spanish textbooks. Also, as shown in Tables 4 and 5, both textbooks present significant amounts of vocabulary that is not present in the frequency dictionary.

The second question posited asks about the under-representation and over-representation of the vocabulary presented. As shown in Table 6, there were not large differences between syntactic categories amongst the textbooks relative to the coverage of those categories in the frequency list. However, *Pido la palabra* (1998) presented about twice as many frequent items as *¡Estoy listo!* (2003). In other words, the under-representation could also be described by comparing the volume of unique lemmas presented from one textbook compared to the other. Also, as seen in Table 7, nouns and adjectives are significantly more likely to be under-represented than verbs and adverbs. This is interesting because even though many more frequent nouns and adjectives are presented, of the vocabulary in these textbooks, a noun or adjective is much more likely to be infrequent than a verb or an adverb that is presented (see Table 5). Because nouns represent an overall much greater number of entries in the frequency list (2,511 out of 5,000), a variable that might obstruct better coverage is the space. There might not be enough room in the materials or time in a single course to incorporate many more concepts.

In terms of over-representation, it was determined (see Table 7) that nouns and adjectives are much more likely to be over-represented in the textbooks than verbs and adverbs. One possible explanation for this is that the semantic fields associated with the chapters' themes might involve more concrete concepts that are easily taught and make for more a more interesting lesson. This is interesting because nouns were also found to be more likely to be under-represented than verbs and adverbs (see Table 6). This means that of the 5,000 most frequent nouns, a smaller percentage is represented in the textbooks. However, nouns were also found to be the most over-represented part of speech, accounting for the majority of infrequent lemmas in both textbooks studied (see Table 7). This is how nouns are able to be both under- and over-represented.

The third question posited asked about potential differences between these SSL textbooks and the SFL textbooks studied by Davies and Face (2006). As discussed in the previous results sections, there were several similarities between Davies and Face's results that were based on SFL textbooks. Like in the Davies and Face study, for example, the textbooks in the current study gave better coverage to higher frequency ranges than lower ones. However, both studies found that the programs that use these textbooks could be significantly improved by including neglected very high frequent lemmas (see section 4.1). Finally, as a small inquiry, it was interestingly found that the second language textbooks examined in the current study were not very variation specific (see Table 8), supporting Moreno de Alba's (2005) claims of the commonality of frequent lemmas across different varieties of a given language.

Overall, the results of this study show that both of the textbooks examined represent the extremely frequent vocabulary well (see Table 7 and Appendix A). However, both also present significant amounts of vocabulary that is not highly frequent, possibly at the

expense of moderately frequent lemmas. Textbooks, including those studied in this investigation, often use themes or situations as a way to create a lesson or chapter. It may not be feasible for a single textbook to cover thousands of frequent words as they appear in a frequency list because of this type of organization. Instead, in a chapter involving foods, several infrequent vocabulary words might be presented because they are conceptually relevant to the lesson. Also, there were some dialectal specific vocabulary entries in both textbooks. However, although these textbooks were designed for students studying in or wanting to study in Mexico, there were much fewer Mexican-specific entries than lemmas common across regional dialects. This might be because there may be little difference between frequent lemmas across different variation; such *mexicanismos* may be seen as more appropriate for more advanced learners; or the textbooks may have been modeled after other, pre-existing resources.

5.0 Conclusions

As concluded by Davies and Face (2006), this research also has shown the potential for corpus linguistics to play a more important role in language teaching, especially in material development. Frequency should play an especially important role in materials designed for second language learners because they need the basic building blocks of the target language in a short amount of time to competently function in the environment in which they live. While there is not as much of a tangible need for first-year, foreign language students to be taught significant percentages of the most frequent words as their second language student counterparts, there does not appear to be any particularly noticeable differences in overall frequency coverage between the SSL textbooks in this study and the SFL textbooks in the Davies and Face study. Authors of future SSL textbooks should thus take into particular consideration presenting as many of the highly frequent lemmas (1-500) as possible (see Appendix A). Even better, a student could be presented with the first 1,000 most frequent lemmas. This would be much closer to the communicative ideal because as Davies (2005) describes, that number of lemmas covers 76-80% of written Spanish and 88% of spoken Spanish. Such coverage, especially for spoken Spanish, would allow a non-native Spanish speaker to be relatively comfortable in an everyday, second language environment.

Because of the quantitative, focused observational nature of this study, second language acquisition as a whole was not well represented. Instead, this study and the methodology it used are a small part of the topic of vocabulary learning and materials development. Further studies would be needed in a more mixed method design to explore how the results from this and similar studies relate to vocabulary learning and teaching as well as textbook design. Mixing these methods with qualitative methods, the researcher

could better understand how beliefs, preferences, attitudes, and other aspects interact with the actual practice and success of vocabulary learning through different approaches and methods.

This thesis is not only an experiment to use a new for of textbook analysis. The results of this study can be applied to real teaching. While there are no judgments made by the researcher, it offers insights into the vocabulary coverage of *¡Estoy listo!* (2003) and *Pido la palabra* (1998). Because these textbooks are so widely used, it could benefit many language program coordinators who currently use or plan on using these textbooks to possibly make decisions of how these textbooks could be implemented best in the design of a particular beginning-level course. Such coordinators and also teachers may never have thought about vocabulary and its relative frequency. Hopefully, this study and others like it will help bring the "vocabulary control movement," which is already strong in ESL and EFL approaches, to the instruction of Spanish.

References

Bade, M. (in press). Grammar and good language learners. In C. Griffiths (Ed.), *Lessons from good language learners* (pp. 146-151). Manuscript submitted for publication.

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

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In Gass, S.M & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 41-68). Cambridge: Cambridge University Press.

Brett, A., Rothlein, L., & Hurley, M. (1996). Vocabulary acquisition from listening to stories and explanations of target words. *The Elementary School Journal*, *96*(4), 415-422.

Canale, M, & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*, 1-47.

Carter, R. (1998). *Vocabulary: Applied Linguistic Perspectives*. London: Routledge.

Carter, R., & McCarthy, M. (Eds.). (1988). *Vocabulary and language teaching*. London: Longman.

Carvajal, C.S., & Horwood, J. (Eds). (2003). *Pocket Oxford Spanish dictionary* (2nd ed.). New York: Oxford University Press.

Chomsky, N. (2000). Internalist explorations. In *New horizons in the study of language and mind*. (pp. 164-194). Cambridge: Cambridge University Press.

Chomsky, N. (1988). *Language and problems of knowledge*. Cambridge, MA: MIT Press.

Collins: The bank of English. (n.d.). Retrieved November 5, 2006, from http://www.collins.co.uk/books.aspx?group=153

Cook, G. (2003). *Applied linguistics*. Oxford: Oxford University Press.

Cortés, M.E. (Ed.) (2003). *¡Estoy listo!: Nivel 1* [I am ready! Level 1] (2nd ed.). Mexico City: UNAM/Santillana.

Davies, M. (2002). *Corpus del Español* [Corpus of Spanish]. Retrieved November 5, 2006, from http://www.corpusdelespanol.org

Davies, M. (2005). Vocabulary range and text coverage: Insights from the forthcoming Routledge frequency dictionary of Spanish. In David Eddington (Ed.), *Selected proceedings of the 7th Hispanic linguistics symposium* (pp. 106-115). Somerville, MA: Cascadilla Proceedings Project.

Davies, M. (2006). *A frequency dictionary of Spanish: Core vocabulary for learners*. London: Routledge.

Davies, M., & Face, T.L. (2006). Vocabulary coverage in Spanish textbooks: How representative is it? In J. Torribio (Ed.), *Selected proceedings from the conference on the acquisition of Spanish and Portuguese in first and second languages*. Somerville, MA: Cascadilla.

Day, R.R., Chenoweth, N.A., Chun, A.E., & Luppescu, S. (1983). Foreign language learning and the treatment of spoken errors. *International Review of Applied Linguistics 5*, 161-170.

Duhne, E.E., Emilsson, E., Montoya, M.T., & del Río, R. (1998). *Pido la palabra: 1er. nivel* [I call for the floor: First level] (7th ed.). Mexico City, Mexico: UNAM.

Ellis, N.C. (1994). Vocabulary acquisition: The implicit ins and outs of explicit cognitive mediation. In N.C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 211-282). London: Academic Press.

Ellis, N.C. (1999). Cognitive approaches to SLA. *Annual Review of Applied Linguistics, 19*, 22-42.

Ellis, N. C. (2001). Reflections on frequency effects in language processing. *SSLA, 24*, 297-339.

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

Gay, L.R., & Airasian, P.W. (2002). *Educational Research: Competencies for analysis and application* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Gordon, R.G., Jr. (Ed.). (2005). Spanish section [Electronic version]. *Ethnologue: languages of the world*, (15th ed.). Dallas, TX: SIL International. Retrieved April 17, 2007, from http://www.ethnologue.com/show_language.asp?code=spa

Griffiths, C. (in press). Age and good language learners. In C. Griffiths (Ed.), *Lessons from good language learners* (pp. 32-39). Manuscript submitted for publication.

Hall, C.J. (2005). *An introduction to language and linguistics: Breaking the language spell*. New York: Continuum.

Ham Chande, R. (1979). Del 1 al 100 en lexicografía [From 1 to 100 in lexicography]. In L.F. Lara (Ed.), *Investigaciones lingüísticas en lexicografía* [Linguistic research in lexicography] (pp. 41-84). Mexico City: Colegio de México.

Hirsh, D., & Nation, I.S.P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a foreign language*, *8*(2), 689-696.

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

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

Krashen S. & Terrell, T. (1983). *The natural approach: Language acquisition in the classroom*. Oxford: Pergamon.

Lara, L.F. (1990). *Dimensiones de la lexicografía: a propósito del diccionario del Español de México* [Dimensions in lexicography: a proposal of the dictionary of Mexican Spanish]. Mexico City: Colegio de México.

Lara, L.F. (1993). *Diccionario fundamental del español de México* [Fundamental dictionary of Mexican Spanish]. Mexico City, Mexico: El Colegio de México.

Lara, L.F. (Ed.). (1996). *Diccionario del español usual en México* [Dictionary of general Mexican Spanish]. Mexico City, Mexico: El Colegio de México.

Larsen-Freeman, D., & Long, M.H. (1991). *An introduction to second language acquisition research*. London: Longman.

Leech, G., Rayson, P., & Wilson, A. (n.d.). *Word frequencies in written and spoken English* (online companion). London: Longman. Retrieved November 29, 2006 from http://www.comp.lancs.ac.uk/computing/research/ucrel/bncfreq/flists.html

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. London: Longman.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove, England: Language Teaching Publications.

Liu, N., & Nation, I.S.P. (1985) Factors affecting guessing vocabulary in context. *RELC Journal 16*(1), 33-43.

Moreno de Alba, J.G. (2005, June). *Unidad y diversidad del español: El léxico* [Unity and diversity in Spanish: The lexicon]. Paper presented at the meeting of clausura del IV curso de la escuela de lexicografía hispánica [closing ceremony of the 4th course of the school of Hispanic lexicography], Madrid, Spain.

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

Nation, I.S.P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*, 35-41.

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

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Real Academia Española. (n.d.). Banco de datos (CREA). *Corpus de referencia del español actual*. Retrieved on November 15, 2006 from http://www.rae.es

Real Academia Española. (n.d.). Proyectos académicos. In *Real Academia Española* (homepage). Retrieved on April 3, 2007 http://www.rae.es/rae/Noticias.nsf/Portada2?ReadForm&menu=2

Renouf, A.J. (1984). Corpus development at Birmingham University. In Aarts, J. & W. Meijs (Eds.), *Corpus linguistics: recent developments in the use of computer corpora in English language research* (pp. 3-39). Amsterdam: Rodopi.

Richards, J.C., & Rodgers, T.S. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge: Cambridge University Press.

Sinclair, J. McH. (1991). *Corpus concordance, collocation*. Oxford: Oxford University

Press.

Sinclair, J. McH., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140-160). London: Longman.

Sökmen, A.J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 237-257). Cambridge: Cambridge University Press.

Stubbs, M. (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, *22*(2), 149-172.

Willis, J.D. (1990). *The lexical syllabus*. London: Collins COBUILD.

Appendix A

Table A1
Lemmas in the Top 500 that were not represented in each textbook

Pido la palabra

	Lemma	Syntactic Category	Frequency	Definition
1	ninguno	adj	144	no, none, nobody
2	cuanto	adj	213	en cuanto a: in terms of, regarding
3	humano	adj	218	human
4	general	adj	227	general
5	político	adj	284	political
6	pobre	adj	373	poor
7	capaz	adj	411	capable, able
8	joven	adj	442	young
9	vivo	adj	453	alive, bright
10	contrario	adj	460	contrary, opposite
11	real	adj	462	royal, real, authentic
12	ambos	adj	488	both
13	principal	adj	496	main, principal
14	tampoco	adv	279	neither, nor, either
15	aún	adv	282	still, yet
16	mal	adv	301	badly
17	mientras	conj	154	while, whereas, as long as
18	embargo	n	180	sin embargo: nevertheless
19	realidad	n	202	reality
20	hecho	n	235	fact, happening
21	fuerza	n	255	strength
22	don	n	303	courtesy title
23	falta	n	344	lack, shortage
24	cara	n	356	face, expression
25	pesar	n	366	sorrow; a pesar de: in spite of
26	ley	n	384	law, bill, rule
27	cuestión	n	398	question, matter
28	partido	n	425	party, group, match
29	derecho	n	427	right, justice, law
30	poder	n	428	power
31	respecto	n	433	respect, con respecto a: with
32	conocimiento	n	434	knowledge
33	resto	n	447	rest, remainder, leftover
34	programa	n	467	program, plan
35	línea	n	473	line, course
36	nivel	n	475	level
37	cabo	n	477	end, bit
38	imagen	n	484	image, picture
39	carrera	n	485	career, course, race
40	figura	n	495	figure
41	contra	prep	172	against, opposite

42	bajo	prep	214	under, underneath
43	ante	prep	236	before, in the presence of
44	cual	pron	153	which, who, whom
45	ello	pron	343	it
46	producir	v	195	to produce, cause
47	permitir	v	220	to allow, permit
48	sacar	v	228	to take out
49	mantener	v	234	to keep, maintain
50	realizar	v	299	to fulfill, carry out
51	comprender	v	306	to understand
52	valer	v	387	to be worth, cost
53	suceder	v	406	to happen
54	dedicar	v	415	to dedicate, devote
55	echar	v	455	to throw, cast
56	obtener	v	466	to obtain
57	soler	v	487	to be accustomed to
58	desarrollar	v	491	to develop

¡Estoy listo!

	Lemma	Syntactic Category	Frequency	Definition
1	alguno	adj	50	some, someone (pron)
2	poco	adj	74	little, few, a little bit (adv)
3	tanto	adj	79	so much, so many
4	tal	adj	120	such (a)
5	mejor	adj	121	best, better (adv)
6	ninguno	adj	144	no, none, nobody
7	solo	adj	160	lonely, alone
8	cuanto	adj	213	en cuanto a: in terms of, regarding
9	humano	adj	218	human
10	igual	adj	239	equal, same (as)
11	distinto	adj	254	distinct, different
12	claro	adj	259	clear
13	cuyo	adj	264	whose
14	bastante	adj	270	rather, fairly, quite a bit (adv)
15	político	adj	284	political
16	demás	adj	312	the rest, others
17	demasiado	adj	335	too much, too many
18	antiguo	adj	348	old, ancient, former
19	pobre	adj	373	poor
20	capaz	adj	411	capable, able
21	natural	adj	414	natural
22	económico	adj	426	economic
23	abierto	adj	439	open, unlocked
24	pasado	adj	445	past, last
25	vivo	adj	453	alive, bright
26	contrario	adj	460	contrary, opposite
27	enorme	adj	471	enormous, vast
28	ambos	adj	488	both

29	profundo	adj	489	deep, profound
30	ya	adv	36	already, still
31	entonces	adv	76	so, then
32	casi	adv	146	almost, nearly
33	nunca	adv	151	never, ever
34	allí	adv	167	there, over there
35	dentro	adv	174	inside
36	ahí	adv	189	there
37	todavía	adv	211	still, yet
38	tampoco	adv	279	neither, nor, either
39	aún	adv	282	still, yet
40	incluso	adv	294	including, even (adv)
41	quizás	adv	297	perhaps, maybe
42	mal	adv	301	badly
43	bueno	adv	337	well . . .
44	través	adv	347	a través: across, over, through
45	pronto	adv	396	soon, quick
46	encima	adv	436	above, on top, in addition
47	fuera	adv	451	out, outside, away
48	lo	art	20	the (+neuter)
49	ni	conj	64	not even, neither, nor
50	pues	conj	103	then, well then
51	sino	conj	109	but, except, rather
52	mientras	conj	154	while, whereas, as long as
53	cosa	n	78	thing
54	hombre	n	80	man, mankind, husband
55	vida	n	88	life
56	forma	n	113	form, shape, way
57	caso	n	130	case, occasion
58	manera	n	152	way, manner
59	tipo	n	157	type, kind
60	gente	n	158	people
61	ejemplo	n	162	example
62	medio	n	171	means, middle; through
63	embargo	n	180	sin embargo: nevertheless
64	modo	n	198	way, manner
65	realidad	n	202	reality
66	obra	n	206	work, book, deed
67	verdad	n	209	truth
68	mes	n	210	month
69	razón	n	212	reason
70	grupo	n	216	group
71	hecho	n	235	fact, happening
72	principio	n	237	beginning, principle
73	pueblo	n	241	people, village
74	fuerza	n	255	strength
75	luz	n	256	light
76	sentido	n	265	sense, feeling
77	paso	n	267	step, pace

78	siglo	n	273	century, age
79	dios	n	274	god, divinity
80	tierra	n	276	earth, land, ground
81	tema	n	283	theme, subject, topic
82	don	n	303	courtesy title
83	final	n	307	al final: finally, in the end
84	fondo	n	318	bottom, end
85	voz	n	320	voice
86	valor	n	326	value, worth
87	necesidad	n	340	necessity, need
88	condición	n	341	condition
89	falta	n	344	lack, shortage
90	estado	n	351	state, condition, status
91	ser	n	352	being
92	cara	n	356	face, expression
93	época	n	358	time, age, period
94	experiencia	n	361	experience
95	pesar	n	366	sorrow; a pesar de: in spite of
96	posibilidad	n	367	possibility
97	resultado	n	379	result, outcome
98	ley	n	384	law, bill, rule
99	aspecto	n	385	aspect, appearance
100	especie	n	388	kind, sort, species
101	cuestión	n	398	question, matter
102	duda	n	399	doubt
103	acción	n	405	action, act, deed
104	peso	n	417	peso (money), weight, load
105	efecto	n	418	effect
106	amor	n	423	love
107	muerte	n	424	death
108	partido	n	425	party, group, match
109	derecho	n	427	right, justice, law
110	poder	n	428	power
111	importancia	n	429	importance
112	suelo	n	432	ground, floor
113	respecto	n	433	respect, con respecto a: with
114	conocimiento	n	434	knowledge
115	libertad	n	435	freedom, liberty
116	esfuerzo	n	444	effort, endeavor
117	resto	n	447	rest, remainder, leftover
118	proceso	n	452	process, procedure
119	nivel	n	475	level
120	gobierno	n	476	government
121	cabo	n	477	end, bit
122	imagen	n	484	image, picture
123	carrera	n	485	career, course, race
124	figura	n	495	figure
125	animal	n	497	animal
126	base	n	498	base, basis

127	hacia	prep	125	toward, towards
128	contra	prep	172	against
129	bajo	prep	214	under, underneath
130	ante	prep	236	before, in the presence of
131	según	prep	257	according to
132	se	pron	9	"reflexive" marker
133	la	pron	33	(direct object)
134	eso	pron	63	that
135	nos	pron	65	us (object)
136	esto	pron	110	this
137	quien	pron	141	who, whom
138	ello	pron	343	it
139	alguien	pron	480	somebody, someone, anyon
140	saber	v	46	to know (a fact), find out
141	parecer	v	81	to seem, look like
142	salir	v	111	to leave, go out
143	volver	v	112	to return, to V again
144	tratar	v	134	to try, treat, deal with
145	existir	v	177	to exist
146	producir	v	195	to produce, cause
147	ocurrir	v	200	to happen, occur
148	entender	v	203	to understand
149	terminar	v	219	to finish, end
150	permitir	v	220	to allow, permit
151	aparecer	v	221	to appear
152	conseguir	v	222	to get, acquire, obtain
153	comenzar	v	223	to begin, start
154	sacar	v	228	to take out
155	mantener	v	234	to keep, maintain
156	resultar	v	238	to result, turn out
157	acabar	v	266	to have just Ved; finish
158	convertir	v	271	to convert, change, become
159	realizar	v	299	to fulfill, carry out
160	suponer	v	305	to suppose, assume
161	comprender	v	306	to understand
162	lograr	v	311	to achieve, get, manage
163	explicar	v	316	to explain
164	reconocer	v	327	to recognize, admit
165	alcanzar	v	329	to reach, catch up with
166	levantar	v	372	to raise, lift
167	intentar	v	376	to try, attempt
168	olvidar	v	383	to forget
169	mostrar	v	392	to show
170	ocupar	v	397	to occupy, use
171	mover	v	402	to move, incite
172	suceder	v	406	to happen
173	fijar	v	407	to se, fix, notice
174	dedicar	v	415	to dedicate, devote
175	aprender	v	422	to learn

176	evitar	v	446	to avoid, prevent
177	interesar	v	448	to interest
178	cerrar	v	454	to close
179	echar	v	455	to throw, cast
180	sufrir	v	457	to suffer, undergo
181	importar	v	464	to matter, import
182	obtener	v	466	to obtain
183	soler	v	487	to be accustomed to
184	desarrollar	v	491	to develop
185	señalar	v	493	to point (out), signal
186	elegir	v	494	to choose, elect
187	proponer	v	500	to propose

Appendix B

Table of Contents of *Pido la palabra: Primer nivel* (1998)

(exerpts, translated)

Classmates, nationalities, professions, time spent in Mexico, activities

Unit 3 An invitation

Thematic content

A telephone call, Activities during one's free time

Vocabulary

Time, shows

Unit 4 Harumi asks for directions and walks through Mexico City

Thematic content

Location

Vocabulary

Places, activities

Unit 5 Looking for lodging

Thematic content

Juan looks for lodging

Vocabulary

Home and its furnishings, colors, numbers

Unit 6 Solving mysteries

Thematic content

Who did it?

Vocabulary

Parts of the body, clothing

Unit 7 Birthday

Thematic content

Birthday party, Plans for a day in the country

Vocabulary

Some colloquial expressions, food and stuff needed for camping

Unit 8 Food

Thematic content

The market, Customs related to food, Preparation of recipes in the kitchen,

In the restaurant

Vocabulary

Food, dishes, kitchen appliances

Unit 9 Daily activities

Thematic content

A normal day

Vocabulary

Daily activities and places

Unit 10 A family album

Thematic content

Family memories

Vocabulary

Family

Unit 11 Tourist places

Thematic content

Plans for a trip, Tourist routes

Vocabulary

Places, baggage, means of transportation

Unit 12 Traditional Mexican parties

Thematic content

Traditional celebrations, Traditions and cultural aspects

Vocabulary

Day of the Dead

Unit 13 The wanderers

Thematic content

Saying good-bye

Vocabulary

(revision)

Appendix C

Methodological Bases of *Pido la palabra: Primer nivel* (1998)

(Summarized and translated from the authors' introduction)

Pido la palabra 1 has a communicative focus, and the acquisition and learning of a language in use are the following:

a) The dialogues attempt to recreate a natural sociolinguistic context, in which what is said is relevant.

b) The characters reflect distinct relationships between interlocutors: friends, acquaintances, family members, loved-ones, etc.

c) The reading material, with a few exceptions, is authentic.

d) The difficulty level of the dialogues in the readings is not completely controlled, with the objective of exposing the student to everyday language in different situations.

e) The book leads the student to develop strategies and to learn inductively, providing him or her with receptive activities and linguistic data with the idea that, over time, he or she will be able to reach his or her own conclusions.

f) The listening comprehension exercises are designed to foment the development of strategies. Understanding everything is not necessary; instead, the goal is to understand the majority of a conversation.

g) The interactive exercises, especially the semi-controlled or free ones, allow the student to use what he or she already knows in new situations and discover what he or she is missing.

h) The book emphasizes oral production; although that is not the ultimate objective.

Appendix D

Table of Contents of *¡Estoy listo!: Nivel 1* (2003)

(excerpts, translated)

Unit 1 In an office

Communicative objectives

To greet

To answer a greeting

To introduce someone/ to introduce oneself

To give personal information

Grammatical objectives

Verb *ser* [to be]: identification, nationality, origin, job, career

Personal pronouns

Definite articles: gender and number agreement with the noun

Adjectives: agreement with the noun

Verb *trabajar* [to work]: example of the first conjugation (present indicative)

Preposition *de* [from]: origin

Interrogative phrases

Lexical objectives

Greetings

Professions, occupations, nationalities

The alphabet

Unit 2 On a trip

Lexical objectives

Days of the week, Numbers (one through 60), cardinal directions, Airport services, Trips, My name is. . .

Unit 3 With the family

Lexical objectives

Numbers (60 through 100), Family, Home, Food

Unit 4 A job interview

Lexical objectives

Hair and skin color, Height and physical complexion, Months of the year, Parts of the human body, Illnesses, Interrogative phrases

Unit 5 Buying and selling

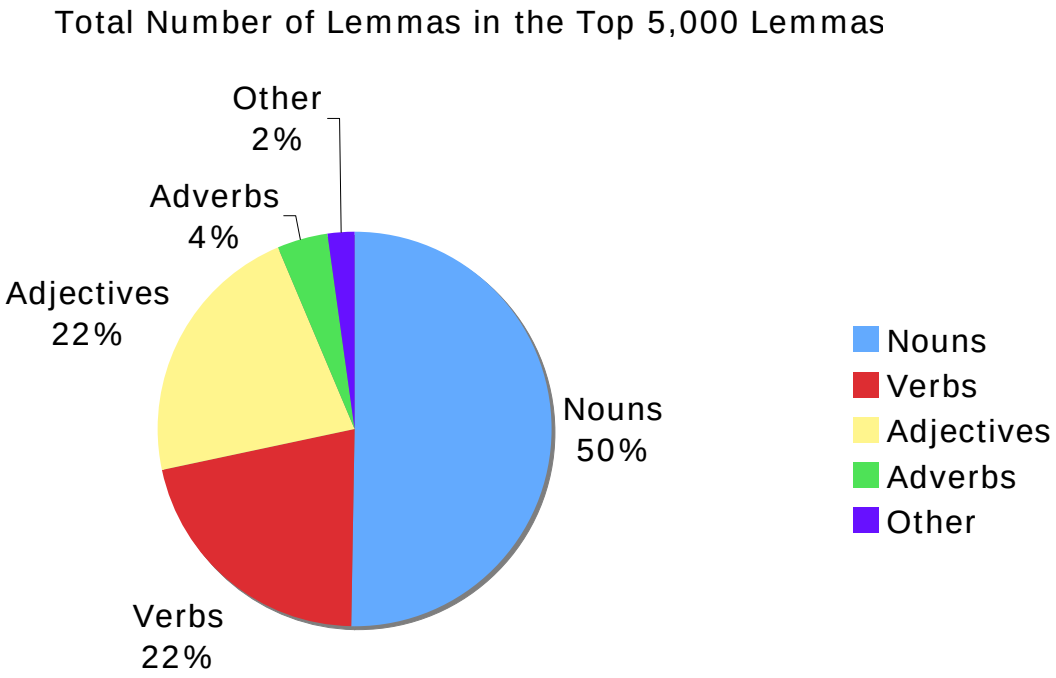
Lexical objectives

Measurements of length and weight, Numbers (100 and above), Clothing and accessories, Colors, Materials, Buying and selling

Appendix E

Figure E1

Percentages of syntactic categories in the top 5,000 most frequent lemmas



Appendix F

Figure F1

Over-represented lemmas