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).