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.