Vocabulary Coverage in Spanish Textbooks: How Representative is It?

Mark Davies and Timothy L. Face
Brigham Young University and University of Minnesota

1. Introduction

Certainly one of the primary goals in developing materials for second language learners should be to create materials that reflect vocabulary and grammar that these learners are likely to encounter in the "real world". There is little to be gained from having students memorize long lists of vocabulary in a textbook, if the learners never again encounter these words once they venture out in the "real world".

The goal of this paper is to determine how well Spanish textbooks do in terms of including realistic, frequency-based vocabulary. We will consider this from two different points of view. First, in quantitative terms, we will consider what percentage of the vocabulary in textbooks appears in standard frequency listings of Spanish vocabulary (and vice versa). Second, in qualitative terms, we will consider what type of vocabulary is both over-represented and also under-represented in Spanish textbooks, vis a vis actual frequency data from other sources.

Before we can address either question, however, we first need to consider two additional questions. First, where can we go to get reliable frequency data for Spanish? Second, how do we measure the vocabulary that appears in textbooks. We will consider the fist two questions in Sections 2 and 3. In Sections 4 and 5 we will then return to the question of representativeness of vocabulary in Spanish textbooks – from both a quantitative and qualitative point of view.

Finally, in a point that we will come back to later in the paper, we should note that this is not a paper on second language acquisition in general, nor the specific issue of vocabulary acquisition. Our focus is strictly the quantitative, corpus-based issue of how vocabulary in Spanish textbooks relates to "real world" language use. Thus this study is only one part of the puzzle, although arguably one of the first pieces needed to create the entire picture.

2. Frequency data

2.1 *The value of frequency-based materials*

In this section we will consider which sources are the most useful for obtaining frequency information on vocabulary use. Some might argue, however, that it is enough to simply rely on "intuition" – that native speakers (at least) would simply "know" which are the most frequent nouns or verbs or adjectives in the language. Yet many researchers have shown that intuition regarding word frequency is often quite inadequate (see Hunston 2002:20-23 for an overview of several different studies). Therefore, it is neither realistic nor fair to ask textbook authors to intuitively "know" which vocabulary should be included in a book. There needs to be a more objective way of determining which vocabulary should be included in textbooks.

Fortunately, corpora can help provide us with such data. To the extent that a corpus accurately represents a wide range of genres and source materials, it can act as a type of "model" for the "real world", and there is now widespread agreement on the general parameters for creating such representative corpora (see Atkins 1993, Kennedy 1998:60-70, Biber et al 1998:246-254). The corpus will contain both spoken and written materials, with a wide range of materials from each genre – for example, conversations, classroom lectures, sermons, unscripted broadcasts, for spoken, and fiction, newspapers, magazines, academic textbooks, email, brochures for written. Some very carefully-crafted corpora such as the 100 million word British National Corpus can serve as useful models of what a highly-representative corpus could and should (and in fact do) look like in terms of different genres and registers (see Burnard 2000).

Cascadilla Proceedings Project
Completed May 4, 2006

Now that representative corpora are available – such as the Longman Corpus (see Biber, et al 1999:1-46), the British National Corpus (Burnard 2000), and the Cobuild corpus (Sinclair 1987), materials developers have begun to use these to create highly useful L2 materials. For example, there are a number of dictionaries that are completely corpus-based, such as the Cambridge Advanced Learner's Dictionary (2005), the Longman Dictionary of Contemporary English (2003), and many others. In terms of grammar books, there are reference works such as the Longman Grammar of Spoken and Written English (Biber et al 1999), which contain detailed statistics on the frequency, distribution, and use of a wide range of grammatical constructions. Finally, there are vocabulary-oriented books for L2 learners which are based entirely on corpus data (e.g. Schmitt and Schmitt 2005).

Before leaving the issue of how corpus-based data relates to textbook vocabulary, we should make one final note. Some might argue that it is unfair to compare the vocabulary of textbooks to "general use" corpora, since textbooks (apparently out of necessity) deal with a reduced set of semantic fields, language functions and discourses. Therefore, the argument might be made that in order to be "fair" in assessing textbook vocabulary, we should create a special corpus that looks just like what a textbook is trying to accomplish – limited semantic fields and language functions. Yet this would miss the mark completely. The goal of a textbook should be to prepare students for the "real world" (as represented by data from an accurate corpus of "real world" speech); the goal should not be that of simply mastering the domains and language functions of "textbook language"..

2.2 *Spanish corpora and Spanish frequency data*

Note that all of the corpora and L2 materials discussed in the previous section are for English. Students of English have benefited for more than a decade now from a wide range of excellent teaching materials that are based on realistic data from highly representative corpora. Unfortunately, for many other languages – such as Spanish – there is either very little or nothing at all that is comparable. For example, a survey of the six textbooks discussed in this paper indicate that none of them claim to have vocabulary based on actual word frequency, and this was the case with several other textbooks that we considered in less detail as well.

The lack of attention to actual word frequency is perhaps due in part to the prevailing culture of textbook publishers for languages such as Spanish. More likely, however, the lack of materials for Spanish has been due to the underlying lack of large, representative corpora for Spanish, on which authors of textbooks could base their vocabulary and grammar. In addition, because there have been few if any large, representative corpora of Spanish until recently, there have likewise been no truly representative frequency dictionaries for Spanish. Without either representative corpora or useful frequency data in dictionaries, materials developers can hardly be faulted for being ignorant of which vocabulary to include in their textbooks. Let us now consider this issue further, and look in more detail at what corpora and frequency dictionaries of Spanish were available until two or three years ago, and how the situation has changed since then.

We will first consider what corpora are available for Spanish, since corpora are the basis of all frequency information. Before 2001, there were no publicly-available corpora of Spanish larger than about one million words. In 2001, the Real Academia Española placed online two large, free corpora. CORDE is a historical corpus (Old Spanish – 1970s) and includes more than 100 million words. CREA is strictly Modern Spanish, and includes more than 120 million words from the 1970s through the current time. The actual text in each corpus is quite impressive, and these corpora have done an admirable job in terms of being representative.

One might hope, then, that the new corpora from the Real Academia could form the database for word frequency of Spanish. Unfortunately, this is not the case. Neither of the two corpora is annotated for part of speech or for lemma (headwords). What this means is that although one could possibly calculate the frequency of individual word forms (*digo, diremos, dijeran*), the lack of lemmatization means that there is no way to group these together under the one lemma *decir*, which is what would need to appear in the vocabulary listing in a textbook. In addition, because the words are not tagged for part of speech, there is no way to know which lemma ambiguous words like *trabajo* belong to (e.g. *trabajo* as noun, or *trabajo* as a form of *trabajar*). Therefore, until very recently, we were still at the point of not having any corpora of Spanish that can be used to create reliable wordlists of Spanish vocabulary.

2.3 *Spanish frequency dictionaries*

Because frequency dictionaries are based on corpora, and because there have not been adequate corpora of Spanish until very recently, it is little surprise that there have not been adequate frequency dictionaries of Spanish either. On the one hand, over time there have been a number of frequency dictionaries, including Buchanan (1927), Eaton (1940), Rodríquez Bou (1952), García Hoz (1953), Juilland and Chang-Rodríguez (1964), Alameda and Cuetos (1995), and Sebastián, Carreiras, and Cuetos (2000).

Yet each of these dictionaries suffers from at least one serious limitation, at least from the point of view of materials developers. First, as can be seen, many of these dictionaries are more than forty years old, and thus would not represent well the Spanish spoken today. Second, all of these frequency dictionaries are based exclusively on written Spanish, and contain no data from the spoken register. In addition, the size of the corpora for all of the dictionaries is quite small – in all cases, less than three million words. The two dictionaries that have been produced in the last ten years both suffer from other important limitations. Alameda and Cuetos (1995) only lists exact forms (e.g. *digo, dices, dijeran*) rather than lemma (e.g. *decir*), and very few of the written texts that it uses are from outside of Spain. The other recent dictionary –Sebastián, Carreiras, and Cuetos (2000) – exists only in electronic form and is extremely hard to acquire, especially outside of Spain.

Among the dictionaries just mentioned, most researchers recognize Chang-Rodríguez (1964) as the most complete frequency dictionary of Spanish to date. Yet because of its methodological limitations (listed above), its list of words is rather skewed, and quite problematic in terms of using the list to create textbook vocabulary. For example, the following words appear high in their list of frequent words, but they would rarely be encountered in the real world by Spanish students (the number in parenthesis indicates the rank frequency order in the dictionary): (nouns) *poeta* (309), *lector* (453), *gloria* (566), *héroe* (601), *marqués* (653), *dama* (696), and *príncipe* (737); (verbs) *acudir* (498), *figurar* (503), *podar* (1932) and *malograr* (2842); (adjectives) *bello* (612), *fecundo* (2376), and *galán* (2557). In addition to including vocabulary that probably would not occur much in the real world, as we will see in a following section, it also omits many highly frequent words that the student would need to know in a real-world setting. Thus, while Chang-Rodríguez (1964) was perhaps the best frequency dictionary until very recently and although it was quite an achievement for its time, it seems clear that forty years later it would not serve as an adequate database for authors of Spanish textbooks.

2.4 *New corpora of Spanish*

Fortunately, the situation regarding Spanish word frequency has changed a great deal in the past four years. First, in 2002 a large, new, publicly-available corpus came online. Second, in late 2005 a frequency dictionary that was based on this corpus was published, which – for the first time – included frequency information that could be used by authors of Spanish textbooks. Let us consider each of these two changes in turn.

In terms of corpora, the *Corpus del Español* came online in late 2002 (www.corpusdelespanol.org). This 100 million word corpus is composed of texts from the 1200s-1900s, with approximately 20 million words from the late 1900s. (This is approximately forty times as large as the corpus used by Chang-Rodríguez.) The corpus is equally balanced between spoken, fiction, and non-fiction, and is also balanced between texts from Spain and those from Latin America. (More details on the composition of the corpus can be found on its website.)

Perhaps the most important aspect of the corpus, vis a vis the corpora from the Real Academia, is that the Corpus del Español is lemmatized and is also tagged for part of speech. This allows us to group together all of the different forms of a word (*digo, diremos, dijeran*) and to assign ambiguous words to the right lemma (e.g. *su cuenta* = *CUENTA*/noun, *que no cuenta* = *CONTAR*/verb). While the tagging and lemmatization that is available via the Corpus del Español website is fairly accurate, this was improved a great deal between 2003-04, and was checked for accuracy in several different ways and at several different points during this time. This has resulted in a corpus that now allows us to find reliable frequency information for Spanish vocabulary.

2.5 *New frequency dictionary of Spanish*

Between 2004-05, the frequency information from the Corpus del Español was packaged together with several other types of information to create *A Frequency Dictionary of Spanish: Core Vocabulary for Learners* (Davies 2005). The dictionary includes basic information like overall frequency of a word in the 20 million word corpus, as well as an indication of range – how well the word is "spread" across all genres in the corpus. In addition, it also includes (for each word) an English gloss, a sample sentence taken from the corpus itself, and an indication of whether the word is much more common or much less common in one register (spoken, fiction, non-fiction) than the others. Finally, there are thirty different thematically-organized lists (foods, clothing, emotions, etc), which show the most frequent words for each of these semantic fields.

One indication of the value of the new frequency dictionary is a comparison of the words that appear in this dictionary, but not in Chang-Rodríguez. These include the following (the numbers in parenthesis show their placement in the new frequency dictionary): (nouns) *oportunidad* 626, *equipo* 737, *película* 827, *control* 889, *televisión* 1079, *rama* 1161, *acceso* 1316, *marca* 1371, *tratamiento* 1419, *experto* 1453, *paciente* 1512, *parque* 1763; (verbs): *enfrentar* 914, *recuperar* 967, *identificar* 988, *controlar* 1071, *transmitir* 1203, *grabar* 1449, *distribuir* 1504, *fallar* 1666, *investigar* 1752, *quebrar* 2376, *apretar* 2405, *fumar* 2472; (adjectives) *capaz* 412, *extraño* 552, *temprano* 1201, *listo* 1457, *ocupado* 1612, *probable* 1842, *latino* 1864, *sucio* 1995, *japonés* 2171, *básico* 2296, *moreno* 2304, *feo* 2382, and *cruel* 2453

In summary, then, the last three or four years have seen the introduction of two important tools – the Corpus del Español and the new *Frequency Dictionary of Spanish: Core Vocabulary for Learners*. With these tools, we have for the first time the frequency information that we need to determine how representative the vocabulary is in different Spanish textbooks – i.e. how well this textbook vocabulary reflects the real world, as measured by a large, representative corpus. Before answering the question of how representative the vocabulary is, however, we must consider one other issue – what criteria we used to extract the vocabulary from the different Spanish textbooks.

3. Selection and compilation of textbook vocabulary for analysis

3.1 *Selection of textbooks*

In order to analyze how representative the vocabulary presented in Spanish language textbooks is of vocabulary encountered in real-word Spanish usage, textbooks had to be selected from which to extract the vocabulary. We decided to include the vocabulary from three first year college-level Spanish textbooks and three second year college-level Spanish textbooks. Since there are a number of textbooks on the market, especially at the first year level, we chose to include textbooks that are in fairly widespread use. Also, at the second year level especially, there are a number of multi-volume texts, with each volume dedicated to a different aspect of the language (e.g. grammar, reading). In spite of the popularity of some of these texts, they were excluded from consideration due to the complications in some cases of each volume having its own set of vocabulary, some of which overlapped between volumes of the text while others did not, and the possibility of instructors using (or at least emphasizing) some volumes and not others. Based on our knowledge of textbooks being used at multiple universities, as well as through consultation with experts in the field of Spanish language textbooks, we selected *Dos mundos*, *Puntos de partida*, and *¿Sabías que…?* as the three first year textbooks and *De paseo*, *Mundo 21*, and *¿Qué te parece?* as the three second year textbooks.

3.2 *Extraction of active vocabulary*

Once the textbooks to be included in the study had been selected, the vocabulary had to be extracted from each textbook. Active vocabulary is the vocabulary that students are expected to learn and be able to use, and is generally the vocabulary included in chapter vocabulary lists. Passive vocabulary, on the other hand, are words that appear in the text, often in reading passages, that 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. Since active vocabulary is the vocabulary that students are expected to learn and be able to use, this best corresponds with what could be expected (at least ideally) to be the vocabulary that students may be prepared to use after taking a course employing a given textbook. Therefore only active vocabulary was extracted from the six textbooks included in the present study. In the glossaries at the end of five of the six textbooks that we examined, active vocabulary was distinguished from passive vocabulary by special notation (for example, indicating the chapter in the text in which each active vocabulary word is presented, but not doing so for passive vocabulary), making extraction of this vocabulary rather simple since all active vocabulary words were contained in one place in the textbook and distinguished from passive vocabulary. In the case of the sixth textbook, examination of the glossary and comparison with chapters in the text made it apparent that only active vocabulary was included in the glossary, and this was confirmed upon consulting the authors of the textbook. Therefore, for all six textbooks active vocabulary was extracted from glossaries at the end of the textbooks, with a separate vocabulary list being compiled for each book.

3.3 *Lemmatization of the vocabulary lists*

As the active vocabulary was extracted from each textbook, the vocabulary list being created was lemmatized. Since a lemmatized vocabulary list includes just one entry for all forms of a given "word", multiple forms listed separately in vocabulary lists were combined into one entry. This is important given that once students know a given lemma, they can generally rather easily create its multiple forms, so a lemmatized list is more representative of vocabulary knowledge than a list that includes many different forms of the same lemma. In addition, the frequency dictionary used for the comparisons in the present study is lemmatized, and thus lemmatization of the vocabulary lists was essential in making comparisons of textbook vocabulary to vocabulary usage in real-world Spanish. To provide two examples of how words in the textbooks were lemmatized, in the case of nouns the singular form (and masculine, when both masculine and feminine forms exist) was entered into the compiled, lemmatized list, and in the case of verbs, the infinitive was entered into the lemmatized vocabulary list. This lemmatization had to be done manually, and meaning was important, as exemplified by words with similar terminations, such as *camarones* and *botones*. The word *camarones* 'shrimp' was presented in textbooks only in the plural form, even though the singular form *camarón* 'shrimp' exists. In this case, in keeping with the structure of the lemmatized vocabulary lists, the singular form *camarón* was entered into the vocabulary list, and in the comparisons run with the frequency dictionary, all forms of this lemma, including *camarón* and *camarones*, were included as occurrences of the lemma represented by *camarón*. With the word *botones*, on the other hand, this form has two separate meanings, where it can be the plural of *botón* 'button' or the singular (or plural) form meaning 'bellhop'. In textbooks, only the meaning of 'bellhop' was included in the vocabulary lists, so the singular form *botones* was included in the lemmatized vocabulary list. Had the word *botones* 'buttons' been presented in the textbooks, this would have been lemmatized as *botón*, the singular form for 'button'.

In creating the lemmatized vocabulary lists, however, several issues arose that are worth mentioning here. First, in some cases separate entries were created for related forms. For example, the verb *almorzar* 'to eat lunch' and the noun *almuerzo* 'lunch' are clearly related, as are the verb *volar* 'to fly' and the noun *vuelo* 'flight'. While these forms are related, they pertain to different parts of speech and therefore are listed as separate lemma in the compiled, lemmatized vocabulary lists. In some cases forms that are listed as separate lemma in the vocabulary list are identical, but again they pertain to different parts of speech. The noun *cerca* 'fence' and adverb *cerca* 'close' as well as the noun *cuarto* 'room' and ordinal number *cuarto* 'fourth' are examples of this. On the other hand, there were multiple forms presented in textbooks that were classified as pertaining to the same lemma, and thus only one entry was made in the vocabulary list. The most notable example of this are the pronouns *él* 'he/him', *ella* 'she/her', *ellos* 'they/them', and *ellas* 'they/them, feminine'. While these forms are presented separately and each given an English gloss, lemmatization results in one form being listed – the masculine singular *él*. In other cases only one out of multiple possible lemma were included in the vocabulary lists. For example, the form *diplomático* can be a noun 'diplomat' or an adjective 'diplomatic', and these would pertain to different lemma. However, only the noun was presented in the text, and so only one lemma for the noun was included in the vocabulary list. Likewise, the form *experto* 'expert' can be a noun or an adjective, but only the lemma for the adjective was included in the vocabulary list since the adjective, but not the noun, was presented in the text. In rarer cases, one of multiple lemma for the same part of speech was included. An example of this is *papa*, which is a noun meaning either ‘potato’ or ‘pope’. Even though these are both nouns, they would have separate entries – as they pertain to separate lemma – if both meanings had been presented in the texts. However, only the meaning of ‘potato’ was presented, and so only one lemma was listed in the vocabulary list. In cases such as those just discussed, a notation was included in the vocabulary list to indicate which lemma was being represented so that this could be correctly coded in running comparisons with the corpus and the frequency dictionary.

Another case that merits mention is the categorization of nouns referring to people, especially as regards both gender and professions. The lemmatization of these forms was in keeping with the design of the corpus and the frequency dictionary with which the vocabulary lists would be compared. In general, similar forms that referred to people but differed in gender were listed as two separate entries in the vocabulary lists. Thus *niño* ‘child or boy’ and *niña* ‘girl’ were listed separately in the vocabulary lists. The exceptions to this were masculine and feminine forms referring to professions, so that *profesor* ‘professor, masculine’ and *profesora* ‘professor, feminine’ were considered two forms of the same lemma *profesor*. In cases where words had masculine and feminine forms that differed beyond the normal masculine-feminine distinctions (and typically these forms correspond to different English words), separate entries were included. Examples of this are the masculine-feminine pairs of *emperador* ‘emperor’ and *emperatriz* ‘empress’, and *actor* ‘actor’ and *actriz* ‘actress’.

Another issue that had to be dealt with were words that were presented as part of an entire phrase, with the meaning of the phrase, but not the individual words, given. An example is *Día de los Reyes Magos* ‘Epiphany’. In this case the entire phrase is glossed by one word. The word *reyes* ‘kings’ was never presented as a separate vocabulary word, so students would never see a gloss indicating the meaning of this word. A similar thing happens with compound words that are written as separate words, such as *arco iris* ‘rainbow’ and *año nuevo* ‘New Year’. In these cases the individual forms making up the compound word or the phrase were each listed separately in the vocabulary list.

Once the vocabulary lists were initially created, they were reviewed and adjusted as necessary to bring them in line with the norms of the corpus and of the frequency dictionary so that accurate comparisons could be made. These comparisons are the basis for the results presented in the remainder of this paper.

4. Quantitative results

To this point, we have discussed two recent tools – the Corpus del Español and the *Frequency Dictionary of Spanish: Core Vocabulary for Learners* – which allow us to adequately judge word frequency in Spanish. We have also discussed how we have been able to extract the vocabulary in several of the most popular textbooks of Spanish. At this point, we will consider – in quantitative terms – how the vocabulary in these textbooks compares with the frequency data from the corpora and frequency dictionary. In other words, how well are these textbooks doing in terms of preparing students for vocabulary that they will encounter in the real word? We will consider three sets of data –coverage by part of speech, by textbook level, by individual textbook

4.1 *Overall coverage and by part of speech*

Our first set of data deal with the overall percentage of words in the textbooks that are found in the frequency dictionary, which is a function of their frequency in the corpus, which presumably is a function of their frequency in the “real world”. The data is summarized in the following table:

Table 1. Vocabulary coverage by frequency and part of speech

Range	Noun	Verb	Adjective	Adverb	Overall
500	0.94	0.92	0.92	0.94	0.93
1000	0.84	0.84	0.83	0.58	0.82
1500	0.72	0.66	0.62	0.41	0.67
2000	0.71	0.64	0.59	0.31	0.66
2500	0.58	0.56	0.50	0.15	0.55
3000	0.55	0.48	0.43	0.19	0.49

3500	0.46	0.33	0.42	0.20	0.42
4000	0.35	0.33	0.39	0.17	0.35
4500	0.39	0.19	0.27	0.07	0.31
5000	0.35	0.22	0.22	0.06	0.29

Table 1 shows the percentage of vocabulary for different ranges in the frequency dictionary, which are included in at least one of the six textbooks. For example, approximately 93% of the words #1-500 in the frequency dictionary (the 500 most common lemma in Spanish) are included in at least one of the six textbooks. This decreases to about 55% of the words in the range 2100-2500, and to about 29% for the words in the range 4501-5000. Intuitively. it should make sense that the percentage of coverage in textbooks would decrease in this way. It would be unlikely that all textbooks would omit highly frequent words like *casa* (word #116), *oír* (263), *correr* (332), or *vivo* (453). On the other hand, it is more likely that less common words would be omitted, like *destreza* (4679), *incidir* (4841), or *nítido* (4908).

Table 1 also shows the coverage by part of speech. As can be seen, the best overall coverage is for nouns, followed by verbs and adjectives (more verbs for words 1-3000, and then adjectives for words 3101-5000), followed by adverbs. Again, this probably makes sense in terms of the goals and competencies of materials developers. Cognitively, nouns are typically more concrete and can be better tied to “real world” objects than verbs and adjectives – thus their overall high occurrence in the textbook vocabulary. On the other hand, adverbs are much harder to connect to real-world objects or events (*acaso, enteramente, inicialmente*). Perhaps more importantly, because many adverbs are derived from adjectives (e.g. *entero* / *enteramente*), if the student has already learned *entero*, it probably is not necessary to include *enteramente* as a separate entry.

4.2 *Coverage as a function of textbook level*

A related question is how vocabulary coverage differs between first and second year textbooks. For example, one might imagine that because first-year textbooks would have (or should have) covered basic vocabulary well (for example, words 1-1000 in the frequency dictionary), these same words might be omitted from second-year textbooks. Conversely, less common vocabulary (e.g. words 3500-5000 in the frequency dictionary), would be found more in the second-year textbooks than those from the first year. What do the data indicate?

Table 2 shows the coverage by year – first-year textbooks vs. second-year textbooks. Again, the data is divided by range in the frequency dictionary – the top 500 words, the next 500 most common words, and so on. In each of the two columns, the figure represents the number of textbooks in each level (a total of three possible in each level), which include each of the words from the frequency dictionary. For example, each of the words #1-500 in the frequency dictionary (*oír, corer, casa*, etc) are – on average – presented in about two (=1.98) of the three first year textbooks. In the second-year books, on the other hand, the same words appear in just about one (=0.95) of the three textbooks.

Table 2. Coverage by level (first-year / second-year textbooks)

Range	First-year	Second-year
500	1.98	0.95
1000	1.24	0.73
1500	0.94	0.49
2000	0.85	0.45
2500	0.63	0.36
3000	0.54	0.36
3500	0.51	0.27
4000	0.31	0.25
4500	0.27	0.20
5000	0.23	0.18

In one sense, the data agrees with our intuitions about the vocabulary focus in the two sets of textbooks. First year books do cover basic vocabulary much better than second-year books. For example, as we have seen, the 500 most frequent words in the frequency dictionary are covered about twice as commonly in the first-year books (=1.98/3.00) than in the second year books (=0.95/3.00).

On the other hand, one might expect that this trend would reverse for the less frequent vocabulary – that this would be covered better in second-year textbooks. However, as the chart shows, there is still more coverage of this vocabulary in the first-year books. Part of this may be due to the fact that many second-year books consciously focus on contemporary issues like "saving the environment" or "fighting against unjust social forces". As admirable as these attempts may be in terms of preparing students to engage in a particular set of social causes, it apparently has the effect of having them focus on vocabulary that – at least outside of those limited domains – will probably be of little value in speaking and writing Spanish.

4.3 *Coverage by individual textbook*

Finally, it might be interesting to consider how consistent the different textbooks are in terms of coverage. Table 3 provides this information. For each textbook, the data show the number of words in the textbook which are found in the 5000 words from the frequency dictionary, the number of words that are not in the frequency dictionary (i.e. they are not one of the 5000 most frequent words in Spanish), and the overall percentage of words in the textbook that are found in the frequency dictionary. For example, *Puntos de Partida* has a total of about 2128 words. 1790 of these words appear in the top 5000 words in the Routledge *Frequency Dictionary of Spanish*, and 328 do not. Therefore, 85% of the vocabulary in *Puntos de Partida* can be found in the frequency dictionary.

Table 3. Coverage by textbook: percentage of words in frequency dictionary

Year	Textbook	+ dictionary	- dictionary	% dictionary
1	*Puntos de partida*	1790	328	0.85
1	*Sabías que*	1306	310	0.81
2	*De paseo*	1580	385	0.80
1	*Dos mundos*	2510	707	0.78
2	*Qué te parece*	396	127	0.76
2	*Mundo 21*	1164	525	0.69

As can be seen, the coverage among the six textbooks is fairly consistent. In terms of the percentage of words in the textbook that are found in the frequency dictionary, the highest figure is 85% for *Puntos de partida* and the lowest is 69% for *Mundo 21*. Again, we see that the first-year textbooks generally have better coverage than the second-year books. Of note is the fact that the total number of words presented in each textbook does vary quite a bit. On the high side, *Dos mundos* presents more than 3200 words in its vocabulary lists, whereas *Qué te parece* include less than one-sixth that amount.

There is another way of looking at vocabulary coverage by textbook, and for some purposes it may even be more insightful than the figures just given. Suppose that a textbook has N number of words, e.g. 1300 words. In the "best of all worlds" scenario, these 1300 words would correspond to words #1-1300 in the frequency dictionary. In other words, it would be as though the textbook vocabulary corresponded exactly to the listing in the dictionary. With this is mind, let us again examine coverage by textbook.

The following table shows the total number of words in each of the six textbooks. It also shows the number of words in the textbook that correspond to words #1-N in the frequency dictionary. For example, there are 3217 words in *Dos Mundos*. 1615 of these 3217 words correspond to words #1-3217 in the frequency dictionary. Therefore, 50% of the words in Dos Mundos correspond to the equivalent range of words in the frequency dictionary.

Table 4. Coverage by textbook: percentage of words in Top N words in frequency dictionary

Year	Textbook	# words overall	# from words 1-N	% of words 1-N
1	*Dos mundos*	3217	1615	0.50
1	*Puntos de partida*	2218	981	0.46
2	*De paseo*	1965	761	0.39
1	*Sabías que*	1616	597	0.37
2	*Mundo 21*	1689	470	0.28
2	*Qué te parece*	523	54	0.10

As can be seen, the highest figure is 50% -- for the textbook *Dos Mundos* – while the lowest figure is 10%, for *Qué te parece*. In other words, 90% of the approximately 520 words in *Qué te parece* do not relate to words 1-520 in the frequency dictionary. As was noted above, this may be a function of the assumed level of the students. The textbook authors may feel that the students are so advanced that they are safe in ignoring "basic" vocabulary and can rather focus on vocabulary related to contemporary social problems. Of course, whether or not this is a valid assumption is open for discussion.

5. Qualitative results: Under- and over-representation in textbooks

5.1 *Under-represented words*

In this section we address words that are under-represented in the Spanish language textbooks that we examined. That is, what words are frequent in real-world Spanish usage as evidenced by the corpus and frequency dictionary, yet are not included in textbooks? In examining under-represented words for this study, we defined under-represented words as those that are among the 1000 most frequent lemma in Spanish but are not included in any of the six textbooks that we included in this study. A breakdown of under-represented words by part of speech reveals that there were 51 nouns (12.8% of all nouns in the top 1000 most frequent words), 34 verbs (12.1%), 28 adjectives (16.9%) and 14 adverbs (19.4%). So while there are more under-represented nouns, this is simply due to the fact that there are more nouns in the top 1000 most frequent words than there are words in the other parts of speech. When percentage of the number of words of each part of speech in the 1000 most frequent words is considered, adverbs prove to be the part of speech that is most under-represented.

Of more interest than the distribution by part of speech, however, is the type of words that tend to be under-represented. The under-represented words tend to represent abstract concepts, and this is true of all parts of speech. Examples of under-represented abstract nouns are *momento* 'moment', *situación* 'situation', *tema* 'topic', *interés* 'interest', and *necesidad* 'need'. Examples of under-represented abstract verbs are *ocurrir* 'to occur', *aparecer* 'to appear', *comenzar* 'to begin', *considerar* 'to consider', and *partir* 'to leave'. Examples of under-represented abstract adjectives are *cualquier* 'any', *único* 'only, unique', *importante* 'important', *especial* 'special', and *diferente* 'different'. When one considers the most concrete of the under-represented words in each category, often they are not nearly as concrete as some other words in that category. For example, among the top few most concrete under-represented words (determined subjectively by the authors), one finds *figura* 'figure', *información* 'information', *distinguir* 'to distinguish', *abandonar* 'to abandon', *ambos* 'both', and *perdido* 'lost'.

Upon examining the example words listed previously in this section, as well as other under-represented words, it is immediately evident that many of the under-represented words are cognates with English words. This raises the question of whether these words might be under-represented simply because they are cognates. Could it be that textbook authors do not include such clear cognates in their active vocabulary lists? While it is important to consider this possibility, it seems fairly clear that this is not the case. Many cognates are, in fact, included in the active vocabulary of all six textbooks that were included in this study. The difference between the included cognates and those that are under-represented is that included cognates tend to be concrete while those excluded tend to be abstract. For example, the following obvious cognates are included in the active vocabulary of the textbooks we examined: *animal* 'animal', *color* 'color', *hotel* 'hotel', *sofá* 'sofa', *organizar* 'to organize', *preparar* 'to prepare', *comentar* 'to comment', *histórico* 'historic', *sexual* 'sexual', and *social* 'social'. So while many of the under-represented words are cognates with English words, it seems clear that they are not under-represented because they are cognates, as numerous clear cognates are included in the active vocabulary of the textbooks. Rather it appears that these words are under-represented because they represent abstract concepts.

5.2 *Over-represented words*

In this section we address words that are over-represented in the Spanish language textbooks that we examined. That is, what words are presented in textbooks in spite of being quite infrequent in real-world Spanish usage as evidenced by the corpus and frequency dictionary? For this study we defined over-represented words as those that are presented in at least two of the six textbooks, and occur 2-100 times in the 20-million word corpus. We ignore words that occur in the corpus one time or not at all, in order to make the numbers manageable, but that is an indication that the results we present here would be even more dramatic if those cases were included. While it might seem like 100 occurrences is a lot, this is not the case when it is considered that this is in a corpus of 20-million words, meaning that the most frequently a word could be used and still be considered over-represented is one occurrence in every 200,000 words of actual Spanish usage. None of these over-represented words occur in the frequency dictionary as one of the 5000 most frequent lemma in the language, indicating just how infrequent these over-represented words are. The 3000 most frequent lemma (and these over-represented words are not in even the top 5000) account for 88-90% of the vocabulary used in written Spanish (depending on whether fiction or non-fiction) and 94% of the vocabulary used in spoken Spanish (Davies 2005:109).

When over-represented words are considered by part of speech, it is clear that nouns are by far the most over-represented, as there are 152 nouns, 19 verbs, 36 adjectives, and no adverbs. This disparity between nouns and the other parts of speech is much more dramatic than in the case of under-represented words, but again it is not simply the distribution by part of speech that is of interest. As with under-represented words, the nature of the words is noteworthy. While under-represented words tend to refer to abstract concepts, over-represented words tend to refer to concrete concepts. Examples of over-represented nouns are *butaca* ‘easy chair’ (2 occurrences in the 20-million word corpus), *calabaza* ‘pumpkin’ (2), *cuchara* ‘spoon’ (2), and *drogadicto* ‘drug addict’ (2). Examples of over-represented verbs are *facturar* ‘to bill’ (13), *freír* ‘to fry’ (21), *patinar* ‘to skate’ (28), and *reciclar* ‘to recycle’ (33). Examples of over-represented adjectives are *aeróbico* ‘aerobic’ (2), *feriado* ‘non-working (day)’ (13), *bilingüe* ‘bilingual’ (39), and *chistoso* ‘funny’ (46). There are very few over-represented words that can be considered even somewhat abstract. In fact, while there are a higher number of adjectives, there are only four over-represented nouns (out of 152) and two verbs (out of 19) that could be considered abstract. The four nouns are *promedio* ‘average’ (5), *ganga* ‘bargain’ (31), *idealista* ‘idealist’ (70), and *noveno* ‘ninth’ (71), though this last one can also be an adjective. The two verbs are *derrochar* ‘to squander’ (40) and *desperdiciar* ‘to waste’ (84). The small number of words that can be considered in any way abstract is an indication of the tendency for over-represented words to represent concrete concepts.

5.3 *The role of semantic fields*

Given that under-represented words tend to represent abstract concepts and over-represented words tend to represent concrete concepts, the question arises as to why that is the case with the vocabulary presented in Spanish language textbooks. We propose that the answer lies in the fact that textbook vocabulary is organized around pre-defined semantic fields. It is undebatable that textbook authors use semantic fields to group vocabulary, reading passages, and other activities. This is evident by looking at the content covered in individual chapters of textbooks, where one often finds such organizing themes as the university, family, shopping, home, weather, environment, food, health and free time. Interestingly, concrete concepts most often fit neatly into these semantic fields, and therefore it is these that are over-represented when authors find themselves needing more vocabulary related to the specific semantic fields included in a textbook. This can be easily exemplified by considering over-represented food and eating terms in the present study. In all, there are more than fifty words related to the topic of food and eating that are over-represented – occurring in two or more textbooks yet not occurring more than once in every 200,000 words of actual Spanish usage. Examples of these over-represented food and eating terms are *calabaza* ‘pumpkin’, *cuchara* ‘spoon’, *galleta* ‘cookie’, *langosta* ‘lobster’, *limón* ‘lemon’, *salsa* ‘sauce’, *freír* ‘to fry’, *tostado* ‘toasted’, *asado* ‘roasted’, *salado* ‘salty’, and *frito* ‘fried’. While these words are quite infrequent in actual Spanish usage, they are concrete concepts that fit the semantic field of food and eating, and thus are over-represented in textbooks as there is a need to add more vocabulary on this topic.

While we argue that concrete concepts are over-represented due to their fit into specific semantic fields that are used in organizing textbooks, we would also argue that under-represented words most often refer to abstract concepts specifically because abstract concepts often do not fit neatly into specific semantic fields. Consider the following examples of under-represented words: *momento* ‘moment’, *situación* ‘situation’, *tema* ‘topic’, *interés* ‘interest’, *necesidad* ‘need’, *ocurrir* ‘to occur’, *aparecer* ‘to appear’, *comenzar* ‘to begin’, *considerar* ‘to consider’, *partir* ‘to leave’, *cualquier* ‘any’, *único* ‘only, unique’, *importante* ‘important’, *especial* ‘special’, and *diferente* ‘different’. While these words are frequent in actual Spanish usage, they do not fit neatly into any of the typical semantic fields used in textbooks and therefore, we argue, they are under-represented. The one notable exception to this pattern are descriptions of people and personalities, which are often abstract and yet do fit into one of the pre-defined semantic fields typical of Spanish language textbooks. Therefore such abstract words as *introvertido* ‘introverted’, *posesivo* ‘possessive’, *impulsivo* ‘impulsive’, *perezoso* ‘lazy’, and *astuto* ‘astute’ are included in Spanish language textbooks, and notably all of these are over-represented words in the present study.

So while it is certainly true that under-represented words tend to represent abstract concepts and over-represented words tend to represent concrete concepts, it is not merely the abstractness or the concreteness of the concept to which a word refers that explains this pattern. Rather the explanation lies in the fact that concrete concepts tend to fit neatly into the types of pre-defined semantic fields around which textbook vocabulary is organized. This again raises the issue set out at the beginning of this paper: one of the primary goals in developing materials for second language learners should be to prepare them for the vocabulary and grammar that they are likely to encounter in the “real world”. The semantic fields around which textbooks are organized certainly serve a purpose in organizing vocabulary, and some might argue that not limiting the corpus to the semantic fields and discourse types found in textbooks makes for an unfair comparison. However, we would argue that the presence of frequency information such as that examined in the present study opens the door for materials developers to move away from the limited semantic fields and discourse types traditionally used in textbooks and bring textbook vocabulary in line with “real world” vocabulary use in the many genres, discourse types and semantic fields that second language learners will encounter.

6. Conclusion

We would like to suggest that corpus linguistics has implications for the development of materials for language teaching. In looking at the results for each textbook in the present study, it stands out that for whatever N number of vocabulary words a textbook includes, only 10-50% of those are among the N most frequent lemma in the language. For example, as Table 4 above indicates, if a textbook presents 2000 vocabulary words, only 10-50% of those words are among the most frequently used 2000 lemma in the language. Yet the most frequent lemma account for a large percentage of actual Spanish usage. If students were to learn only the most frequent 1000 lemma in Spanish, they would be fairly well prepared to understand much of the Spanish vocabulary they would encounter in the real world, as the 1000 most frequent lemma account for 76-80% of written Spanish (depending on whether fiction or non-fiction) and 88% of spoken Spanish (Davies 2005).

While pre-defined semantic fields are certainly a convenient way to organize the vocabulary presented in textbooks, it appears that this type of organization leads to words that refer to concrete concepts being over-represented and words that refer to abstract concepts being under-represented. While we certainly do not wish to suggest that semantic fields be abandoned in textbook design, we suggest that it would be useful for materials designers to find ways to take word frequency into account in determining the vocabulary to be included in Spanish language textbooks.

We should also note how our findings might be integrated in with other research in the future. This study has not dealt with second language acquisition per se, or the specific issues of vocabulary acquisition and the relationship between input and proficiency. Our orientation has been strictly that of a quantitative, corpus-based perspective, and it is just one piece of the puzzle – although arguably one of the first pieces needed to address the issue. We now invite other SLA researchers to integrate these corpus-based findings into their own research, to look comprehensively at the big picture of "real world" language use, vocabulary acquisition, and materials development.

In summary, then, we have seen that corpora have been instrumental in helping to create realistic second-language materials for other languages such as English. This is due in large part to the accurate and representative corpora and the derived frequency listings that are available for these languages. In the case of Spanish, on the other hand, this frequency material has become available only very recently. With this data we can now begin to look more critically at the material presented in Spanish textbooks, in much the same way that this has been carried out for other languages for several years now. Our hope is therefore that this study might be the first step in helping to create more realistic and useful textbooks for learners of Spanish.

Textbooks used in the study

Knorre, Marty, Thalia Dorwick, Ana María Pérez-Gironés, William R. Glass and Hildebrando Villarreal. 2001. *Puntos de partida*, 6th edition. Boston: McGraw-Hill.

Lee, James F., Dolly Jesusita Young, Rodney Bransdorfer and Darlene L. Wolf. 2005. ¿*Qué te parece*?, 3rd edition. Boston: McGraw-Hill.

Long, Donna Reseigh and Janice Lynn Macián. 2005. *De paseo*, 3rd edition. Boston: Heinle.

Samaniego, Fabián A., Nelson Rojas, Maricarmen Ohara and Francisco X. Alarcón. 2004. *Mundo 21*, 3rd edition. Boston: Houghton Mifflin.

Terrell, Tracy D., Magdalena Andrade, Jeanne Egasse and Elías Miguel Muñoz. 2002. *Dos mundos*, 5th edition. Boston: McGraw-Hill.

VanPatten, Bill, James F. Lee and Terry L. Ballman. 2000. ¿*Sabías que*…?, 3rd edition. Boston: McGraw-Hill.

References

Alameda, J.R. & Cuetos, F. 1995. *Diccionario de Frecuencias de las Unidades Lingüísticas del Castellano*, Oviedo: Universidad de Oviedo.

Atkins, et al (1993) "Corpus design criteria". *Literary and Linguistic Computing* 7:1-16.

Biber, Douglas, et al. 1998 *Corpus linguistics: Investigating language structure and use*. Cambridge UP.

Biber, Douglas, et al. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Buchanan, M.A. 1927. *A Graded Spanish Word Book*. Toronto: Univ. of Toronto Press.

Burnard, L. 2000. *Reference Guide for the British National Corpus* (World Edition). Oxford: Oxford University Computing Services.

Cambridge Advanced Learner's Dictionary. 2005. 2nd edition. Cambridge: Cambridge UP.

Davies, M. 2002. *Corpus del Español*. Free online access at http://www.corpusdelespanol.org

-------. 2005. *A Frequency Dictionary of Spanish: Core Vocabulary for Learners*. London: Routledge.

-------. 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*, 106-115. Somerville, MA: Cascadilla Proceedings Project.

Eaton, H. (1940) *An English -French - German - Spanish Word Frequency Dictionary*. New York: Dover Publications.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge UP.

García Hoz, V. 1953. *Vocabulario Usual, Vocabulario Común y Vocabulario Fundamental*. Madrid: CSIC.

Juilland, A. & Chang-Rodríguez, E. 1964. *Frequency Dictionary of Spanish Words*. The Hague: Mouton.

Longman Dictionary of Contemporary English. 2003. New York: Longman.

Kennedy, Graeme 1998. *Introduction to Corpus Linguistics*. Longman. 60-70.

Sinclair, J.M., ed. 1987. *Looking Up: an Account of the COBUILD Project in Lexical Computing*. London: Collins.

Rodriguez Bou, L. 1952. *Recuento de Vocabulario Español.* Rió Piedras : Universidad de Puerto Rico.

Schmitt, Diane, and Norvert Schmitt. 2005. *Focus on Vocabulary: Mastering the Academic Word List*. New York: Pearson ESL

Sebastián, N., Martí, M.A., Carreiras, M.F. & Cuetos, F. 2000. *LEXESP, Léxico Informatizado del Español*. Barcelona: Ediciones de la Universitat de Barcelona. (CD-ROM only)