Vocabulary Range and Text Coverage: Insights from the Forthcoming *Routledge Frequency Dictionary of Spanish*

Mark Davies
Brigham Young University

1. Introduction

An important question for natural language researchers, general linguists, and even teachers and students is how much text coverage can be achieved with a certain number of lexemes in a given language. In studies such as National (2000), we find that the top 1000 lexemes in English account for about 80% of all tokens in a given text. The second block of 1000 lexemes provides coverage for approximately 5% additional coverage of tokens, and this drops to about 3-4% for the third set of 1000 lexemes. These data are important for language learners (and teachers), as they attempt to address the issue of core vocabulary, and how much time and effort should be spent in extending vocabulary beyond a certain level.

While studies of vocabulary coverage have been carried out for other languages (see, for example, Jones 2003), none has been carried out for Spanish. Most likely, the reason for this is that until very recently, we did not have the raw materials upon which to base such a study. In order to provide an accurate model of the Spanish lexicon, we must first have a representative corpus, including texts and transcripts of conversation from a wide variety of genres and registers. These texts must then be accurately annotated for part of speech and lemma. The present study is an overview of how this process has been carried out in the creation of the Frequency Dictionary of Spanish, which will be published by Routledge in 2005.

2. Previous studies of vocabulary frequency in Spanish

There have already been a number of frequency dictionaries of Spanish, and one might suppose that the data from one or several of these would be sufficient to study text coverage with a given level of vocabulary in Spanish. Unfortunately, this is not the case. The most accurate frequency study of Spanish to date is probably Chang-Rodríguez (1964). While it was a notable achievement for its time, it has become somewhat outdated since that time. The corpus on which the frequency data is based is only one million words, and all of it comes from strictly literary works, and solely from peninsular texts.

Because there is no spoken component to the corpus, the vocabulary is highly skewed. For example, the word *poeta* is word number 309 in the frequency list, with other cases like *lector* (453), *gloria* (566), *héroe* (601), *marqués* (653), *dama* (696), and *príncipe* (737). This skewing is not limited just to nouns, but also includes what would in a normal corpus be much lower frequency verbs, like *acudir* (number 498 in the complete frequency list), *figurar* (503), and *juzgar* (560) and adjectives like *décimo* (240) and *bello* (612). Again, the skewing is due more to the inadequate corpus on which the frequency list is based rather than being a product of the general methodology, and it is simply a function of the difficulty in creating large, representative corpora forty years ago. Such is also the case with the Brown corpus of American English, which -- like the corpus used for Chang-Rodríguez -- was based on just one million words from strictly written texts -- and yet which nonetheless remained the standard corpus of English for more than thirty years.

In addition to Juilland and Chang-Rodríguez (1964), there have been a number of other frequency dictionaries and lists for Spanish (Buchanan 1927, Eaton 1940, Rodríquez Bou 1952, García Hoz 1953, Alameda and Cuetos 1995, Sebastián, Carreiras, and Cuetos 2000), but all of these suffer from significant limitations as well. Most importantly, all of the frequency dictionaries are based exclusively on written Spanish, and contain no data from the spoken register. This leads to the type of unrepresentative vocabulary shown above. In addition, five of the dictionaries (Buchanan 1927, Eaton 1940, Rodríquez Bou 1952, García Hoz 1953, Juilland and Chang-Rodríquez 1964) are now quite outdated and are based on texts from the 1950s or earlier. In addition to being based strictly on written Spanish, 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 – rather than lemma – and very few of the written texts on which it is based are from outside of Spain. Finally, Sebastián, Carreiras, and Cuetos (2000) exists only in electronic form and is extremely hard to acquire, since it can only be purchased (at least at the present time) directly from the University of Barcelona.

3. Corpus and methodology

The goal, then, has been to create a representative corpus of Spanish, annotate it for part of speech and lemma, and then use this data to examine lexical coverage with varying levels of lexemes (top 1000 words, top 2000 words, etc). The corpus used for this study is fairly similar to the sub-corpus from the 1900s that is found in the 100 million word NEH-funded Corpus del Español that I completed in 2002 (see Davies 2002; also http://www.corpusdelespanol.org).

The Corpus del Español is highly annotated, and thus allows a wider range of searches than almost any other large corpus in existence. For example, users can search by part of speech and lemma (1-2), wildcards (3), synonyms (4), and customized lists (5). In addition, all queries are very fast. Due to the unique database architecture that I have created, even the most complex queries take only 3-4 seconds to search the entire 100 million words.

Table 1. Types of searches in the Corpus del Español

1	*.pn_obj querer.* *.v_inf	lo quiero hacer, me quería hablar
2	*.n suave.*	voz suave, viento suave, inviernos suaves
3	s_fr_r*	sufrir, sofre, sufrirán
	tan * como	tan bueno / bien / grande como
4	!difícil.* de *.v_inf	difícil de hacer, imposible de evitar
5	[se] poner.* el/la [lópez:ropa].*	[se] puso la chaqueta, [se] pone el sombrero

In 2002 Doug Biber of Northern Arizona University and I received a two year grant from the National Science Foundation to use the Corpus del Español as part of a "multidimensional analysis" of Spanish (see Biber and Davies 2002). As part of this project, we have modified the corpus somewhat, to make sure that all of the texts are from the 1970s or later, and that there is nearly equal distribution of texts from spoken Spanish, written/fiction, and written/non-fiction. The following table provides an overview of the genres and registers in the 20 million word corpus. Note that the size refers to millions of words, and the superscript numbers refer to the sources following the table.

Table 2. Composition of 20 million word Modern Spanish corpus

	# words	Spain	# words	Latin America
Spoken	1.00	España Oral[1]	2.00	Habla Culta (ten countries)
	0.35	Habla Culta (Madrid, Sevilla)		
3.35	**1.35**		**2.00**	
Transcripts/ Plays	1.00	Transcripts/Interviews (congresses, press conferences, other)	1.00	Transcripts/Interviews (congresses, press conferences, other)
	0.27	Interviews in the newspaper ABC		
	0.40	Plays	0.73	Plays

3.40	**1.67**		**1.73**	
Literature	0.06	Novels (BV[2])	1.60	Novels (BV[2])
	0.00	Short stories (BV[2])	0.87	Short stories (BV[2])
	0.19	Three novels (BYU[3])	1.11	Twelve novels (BYU[3])
	2.17	Mostly novels, from LEXESP[4]	0.18	Four novels from Argentina[5]
			0.20	Three novels from Chile[6]
6.38	**2.42**		**3.96**	
Texts	1.05	Newspaper ABC	3.00	Newspapers from six different countries
	0.15	Essays in LEXESP[4]	0.07	Cartas ("letters") from Argentina[5]
	2.00	Encarta encyclopedia	0.30	Humanistic texts (e.g. philosophy, history from Argentina[5)]
			0.30	Humanistic texts (e.g. philosophy, history from Chile[6)]
6.87	**3.20**		**3.67**	
Total	**8.64**		**11.36**	

Sources:
1. *Corpus oral de referencia…* (http://elvira.lllf.uam.es/docs_es/corpus/corpus.html)
2. The *Biblioteca Virtual* (http://www.cervantesvirtual.com)
3. Fifteen recent novels, acquired in electronic form from the Humanities Research Center, Brigham Young University
4. *Léxico informatizado del español* (http://www.edicionsub.com/coleccion.asp ?coleccion=90)
5. From the C*orpus lingüístico de referencia de la lengua española en argentina* (http://www.lllf.uam.es/~fmarcos/informes/corpus/coarginl.html)
6. From the C*orpus lingüístico de referencia de la lengua española en chile* (http://www.lllf.uam.es/~fmarcos/informes/corpus/cochile.html)

Once the corpus was created, we then created a tagger to assign part of speech and lemma information to each form. At the most basic level, this is necessary to group together all forms of a verb or noun, since only the headword (e.g. *decir*) will appear in the frequency dictionary, containing the aggregate for all of the different verbal forms (*dice, dijeron, diremos*, etc). The problem, of course, comes with forms that belong to more than one lemma, such as *fue* (*ser*/*ir*), or *sienta* (*sentar, sentir*), or which belong to more than one part of speech, such as *trabajo* (*trabajar*/*el trabajo*), *como* (*comer*/*como*), and *casa* (*casar*/*la casa*). In all of these cases, the tagger uses certain heuristic devices and algorithms to correctly disambiguate the forms, so that each word form in the corpus ends up being categorized into the correct lexeme. Obviously, this is a non-trivial task, and is based on general strategies of disambiguation that have been applied to many other languages (see, for example, Garside et al, 1997).

4. Vocabulary coverage

With the frequency data from the annotated corpus, we were then able to extract lists of the 6000 most frequent lexemes, which will form the basis of the Routledge Frequency Dictionary of Spanish. However, we can also use this same data to examine the issue of text coverage with differing levels of lexemes, which is the focus of this paper. In the following table -- which represents the main conclusions of this study -- we see the percent coverage of all tokens in three different registers (oral, fiction, and non-fiction) at three different levels of lexemes -- top 1000 words, top 2000 and top 3000.

Table 3. Percent coverage of tokens by groups of types/lemma

	Non-fiction	Fiction	Oral
1st thousand	76.0	79.6	87.8
2nd thousand	8.0	6.5	4.9
3rd thousand	4.2	3.5	2.3
FIRST 3000	88.2	89.6	94.0

As the data indicate, a limited vocabulary of 1000 words would allow language learners to recognize between 75-80% of all lexemes in written Spanish, and about 88% of all lexemes in spoken Spanish (which is due to the higher repetition of basic words in the spoken register). Subsequent extensions of the base vocabulary have increasingly marginal importance. By doubling the vocabulary list to 2000 words, we account for only about 5-8% more words in a given text, and the third thousand words in the list increases this only about 2-4% more. There clearly is a law of “diminishing returns” in terms of vocabulary learning.

The following table indicates how the data from Spanish compares to that of Nation (2000) for English and Jones (2003) for German.

Table 4. Percent coverage of tokens in different languages

	Non-fiction			Fiction			Oral		
	Span	Eng	Ger	Span	Eng	Ger	Span	Eng	Ger
1st thousand	76.0	74.5	64.7	79.6	82.3	72.0	87.8	84.3	82.6
2nd thousand	8.0	4.7	7.2	6.5	5.1	5.4	4.9	6.0	4.4
FIRST 2000	84.0	79.2	71.9	86.1	87.4	77.4	92.7	90.3	87.0

The data from Spanish and English are roughly comparable, but there is an important difference in the way in which the data was obtained. In Nation (2000), the words are grouped by what he calls “word families”, so that [*courage, discouragement, encourage*] would all be grouped under the headword [COURAGE], and [*paint, painted, painter, painting*] would all be grouped under the headword [PAINT]. In our study, however, we used the traditional lemma approach, in which *pintar, pintura, pintor, and pintoresco* would all be assigned to different lemma, and [*pintamos, pinto,* and *pintarás*] would all be assigned to the lemma [PINTAR]. Because we separate the nominal, verbal, and adjectival uses, we might expect that the same number of headwords would lead to less text coverage than in English. The fact that this does not happen, however, is probably due to the fact that English has a larger lexical stock than Spanish, due to the influence of native Anglo-Saxon and imported Franco-Norman and Latinate words (e.g. *real, royal, regal*). The fact that the same amount of lexemes in German leads to lower textual coverage is somewhat more difficult to explain. It may be due to the still-incomplete state of the German tagger (Jones, p.c.). Or again, it may be due to a generally larger lexical stock in German than in Spanish, though this is much more debatable.

5. Text coverage as a function of part of speech and genre

As seen in Tables 3-4 above, the degree of textual coverage is in part a function of the register. 1000 words will provide more coverage in spoken Spanish than they will in written Spanish, because a typical written text uses a wider range of vocabulary than a typical conversation. It is also a function, however, of the part of speech, as is seen in the following tables.

Table 5 shows what percent of token coverage is obtained by taking the most frequent lemma in a particular register. For example, there are about 12900 distinct nouns in the register of spoken Spanish (the figure at the bottom of the second column from the left). Twenty percent of these 12900 nouns would be about 2578 nouns. If one takes the most frequent twenty percent of the nouns (the cell indicated with bolding), this will account for about 248,000 tokens, or about 89% of all nouns. By doubling the amount of nouns (i.e. 40% or 5155 types), one obtains another 6% of all of the tokens (of nouns) in the corpus.

Table 5. Vocabulary coverage in spoken texts, by percent of lemma

		N			V			ADJ			ADV		
2%	258	50.4	140874	47	74.9	207378	66	43.4	24607	12	65.6	98467	3851
5%	645	67.5	188587	116	84.7	234407	165	61.6	34885	29	83.9	125910	
10%	1289	79.3	221496	231	90.6	250911	329	75.1	42537	57	93.1	139599	
20%	2578	88.8	248045	462	98.6	272908	657	85.9	48645	114	97.8	146725	
30%	3867	93.0	259611	693	99.1	274421	985	90.9	51493	171	99.0	148521	
40%	5155	95.3	266233	923	98.6	272908	1314	93.9	53192	227	99.4	149168	
50%	6444	96.9	270563	1154	99.1	274421	1642	95.9	54306	284	99.7	149497	
60%	7733	97.9	273355	1385	99.5	275346	1970	97.2	55075	341	99.8	149686	
70%	9021	98.6	275403	1615	99.7	275958	2299	98.3	55667	397	99.9	149804	
80%	10310	99.1	276692	1846	99.8	276359	2627	98.8	55995	454	99.9	149887	
90%	11599	99.5	277981	2077	99.9	276590	2955	99.4	56323	511	100.0	149944	
	12897		279269	2314		276820	3293		56651	568		150000	

As can be seen, the degree of coverage is a function of the part of speech. Assume that a language learner is aiming for 90% coverage in each of the four parts of speech that represent open classes -- nouns, verbs, adjectives, and adverbs. This 90% figure will be obtained by knowing about 2600 nouns, 230 verbs, 980 adjectives, and 50 adverbs, or a total of about 3800 total forms. This is important in terms of the number of words that one would want to have in a frequency dictionary. If one adds the figure of about 3800 words to a list containing the more basic function words (determiners, prepositions, conjunctions, etc), it suggests that a frequency list with about 4000 words total would cover about 90% of all words that would be heard in a typical conversation. By increasing this to about 6000 words, one would account for about 3-4% more words in a typical conversation.

Of course lexical coverage is a function of register as well. In a technical and highly-specialized piece of non-fiction writing, for example, one might expect the same list of 4000 words to provide a lower degree of coverage than in the spoken register. In fact this is the case, as can be seen in the following two tables, which show coverage in fiction and non-fiction writing. As Table 6 shows, one would need about 7000 words to achieve 90% coverage in fiction writing, and nearly 8000 words for non-fiction. The Routledge Frequency Dictionary of Spanish will have approximately 6000 words, which will translate to about 85-90% coverage of a typical text, depending on the register.

Table 6. Vocabulary coverage in fiction texts, by percent of lemma

		N			V			ADJ			ADV		
2%	314	41.4	112961	61	57.1	117649	83	37.8	18954	15	61.2	39286	7043
5%	784	58.1	158489	152	70.5	145228	208	53.4	26764	37	83.0	53317	
10%	1568	71.6	195378	304	80.8	166441	415	66.1	33104	74	93.7	60213	
20%	3135	84.1	229326	607	89.8	184926	830	78.8	39484	148	97.0	62306	
30%	4702	90.1	245845	910	94.1	193715	1245	85.9	43021	221	98.0	62977	
40%	6269	93.6	255389	1214	96.4	198529	1660	90.4	45280	295	98.7	63384	
50%	7836	95.8	261345	1517	97.8	201383	2074	93.5	46822	368	99.1	63650	
60%	9404	97.2	265221	1820	98.7	203226	2489	95.7	47931	442	99.4	63842	
70%	10971	98.3	268022	2124	99.3	204399	2904	97.3	48761	516	99.6	63990	
80%	12538	98.9	269589	2427	99.6	205139	3319	98.3	49270	589	99.8	64084	
90%	14105	99.4	271156	2730	99.9	205580	3734	99.2	49685	663	99.9	64158	
			272723			205883			50099			64231	

Table 7. Vocabulary coverage in non-fiction texts, by percent of lemma

		N			V			ADJ			ADV		
2%	601	52.4	319013	51	47.9	108047	87	41.0	55303	12	63.3	37206	7870
5%	1502	70.8	431095	127	63.7	143597	217	58.2	78474	29	79.6	46794	
10%	3003	82.6	502863	254	77.2	174197	433	72.6	97974	58	88.8	52177	
20%	6005	90.9	553108	508	89.6	202030	866	85.4	115171	115	94.8	55727	
30%	9008	94.2	573372	762	94.6	213411	1299	91.2	123071	173	97.2	57101	
40%	12010	96.0	584484	1015	97.1	218937	1732	94.6	127604	230	98.3	57756	
50%	15013	97.2	591460	1269	98.3	221837	2164	96.7	130394	288	98.9	58138	
60%	18015	98.0	596600	1523	99.1	223477	2597	97.9	132140	345	99.3	58386	
70%	21018	98.5	599603	1776	99.5	224430	3030	98.8	133274	403	99.6	58551	
80%	24020	99.0	602605	2030	99.7	225004	3463	99.4	134043	460	99.8	58655	
90%	27023	99.5	605608	2284	99.9	225316	3896	99.7	134476	518	99.9	58713	
			608610			225569			134908			58770	

6. The importance of range

There is one final consideration for words that should be included in a “basic Spanish” frequency list, and that is range. There are cases in which a word has high overall frequency, but it is not distributed evenly throughout the corpus. For example, it may occur in just one or two specialized articles, and then not be found at all in any of the other blocks of text. The following tables provide some examples of this phenomenon. The words in both tables have roughly the same frequency -- about 50-70 occurrences per million words of text. The words in the table to the left, however, have a wide range in the corpus, and occur in at least 60 of the 100 blocks of text. (The blocks were created by simply dividing the 20 million words into 100 equally-sized blocks of 200,000 words each.) The words in the table to the right, on the other hand, occur in less than 15/100 blocks of text. The contrast between the two is quite striking.

Table 8. Basic difference between vocabulary with wide/narrow range

range > 60 (/100)		
range		freq
68	notable	66
67	falta	63
66	introducción	69
65	preocupación	65
64	propósito	69
64	disposición	69
64	empleo	68
64	peligro	67
64	difusion	62
64	duda	57
62	protección	69
62	clave	60
62	reconocimiento	59
62	precedente	58
62	complete	54
61	impacto	55
61	margen	53

range <= 15 (/100)		
range		freq
7	verbo	69
7	cromosoma	58
7	neutrón	56
9	bailarín	65
9	sonata	62
10	cirugía	56
11	galaxia	54
12	enciclopedia	54
13	glándula	61
13	fármaco	59
13	filo	58
14	orquesta	58
14	jazz	57
14	corán	53
15	turbina	61
15	enzima	54

Although the overall frequency of the words in the two lists is the same, one intuitively senses that wide-range words like *falta, propósito, peligro*, and *duda* probably should be in the frequency dictionary, whereas narrow-range words like *cromosoma, sonata, galaxia*, and *glándula* probably should not. In the case of the *Routledge Frequency Dictionary of Spanish*, we have set a lower limit of 20 for range, to make sure that the words contained therein are part of everyday Spanish, rather than an overly-narrow technical domain.

7. Range, frequency, and genre

In the preceding section we suggested that words that do not occur above a certain threshold -- in terms of range -- should probably not be included in the frequency dictionary. However, what about a case where the word meets that threshold -- in terms of the entire corpus -- but has limited range in a particular register? For example, the following table contains words that have wide range in the spoken register, but limited range in non-fiction writing (e.g. 97 range with *preguntar* in spoken, 12 in written, for a difference of 85). There are hundreds of such words; in this table we have simply chosen the ten words with the largest difference in range.

Table 9. Wide range in oral, limited range in non-fiction

range			verb	frequency - per million words		
diff	oral	non-fiction		diff	oral	non-fiction
90	99	9	gustar	1296	1301	5
85	97	12	preguntar	266	272	6
82	88	6	meter	315	318	3
75	94	19	imaginar	247	258	11
68	71	3	encantar	161	162	1
65	86	21	casar	309	324	15
64	96	32	tocar	296	356	60
62	77	15	echar	177	184	7
62	85	23	faltar	140	157	17
62	89	27	comprar	331	355	24

Likewise, Table 10 shows words that have wider range (and frequency) in non-fiction writing than in the spoken register.

Table 10. Wide range in non-fiction, limited range in spoken

range			verb	frequency - per million words		
diff	non-fiction	oral		diff	non-fiction	oral
87	100	13	denominar	455	468	13
82	93	11	contener	314	326	12
80	86	6	añadir	135	142	7
80	97	17	situar	326	346	20
77	88	11	sustituir	148	160	12
76	87	11	combiner	163	177	14
76	92	16	proporcionar	229	239	10
73	97	24	componer	209	240	31
73	93	20	contribuir	110	140	30
73	92	19	adopter	282	308	26

Finally, while it is probably not surprising that there is a difference between the spoken and non-fiction registers, one might suppose that there would be much less difference between the spoken register and fiction writing. But even here, there are some noticeable differences, as seen in the following table.

Table 11. Wide range in fiction, limited range in spoken

range			verb	frequency - per million words		
diff	fiction	oral		diff	fiction	oral
78	84	6	sonreír	281	288	7
76	78	2	murmurar	232	233	1
70	90	20	soltar	210	241	31
70	93	23	gritar	396	420	24
69	72	3	acariciar	182	184	2
68	75	7	besar	200	206	6
68	75	7	oler	157	163	6
67	81	14	encender	204	216	12
66	78	12	asomar	139	148	9
66	78	12	envolver	135	143	8

These three tables represent cases, then, in which the word has the requisite overall range to be in the dictionary, but there is still a large difference in range (and usually frequency as well) from one register to another. In these cases, there will simply be a short annotation in the dictionary entry (e.g. S [spoken], F [fiction], or NF [non-fiction]) to indicate that the word is found much more commonly in a subset of the registers. This will hopefully be of value to language learners as they attempt to understand the actual distribution of a given word.

8. The organization of the frequency dictionary

The preceding discussion helps to explain the criteria that are used to determine which words should be in the frequency dictionary. In this section we briefly provide some sense of the final organization of the dictionary, and the type of information that will be found in each index. The primary index will contain the entries in rank frequency order, starting with the most common word (*de*), and continuing through the top 6000 lemma of Spanish. Although the final list has not yet been completely determined, the following table presents sample entries for some words that will appear at about frequency number 1500. As can be seen, each entry contains 1) the rank frequency order 2) part of speech 3) English gloss 4) actual example from the 20 million word corpus 5) frequency count (number of occurrences), and (if applicable) a notation indicating in which register the word is more frequent, if there is a significant difference (note that this particular feature is given just for the purposes of illustration in this table, and does not represent actual frequency distribution).

Table 12. Frequency listing in the dictionary

1500 **asociación** *n* ‘association’ *en estos países no existen las asociaciones de socorro* 1199	1509 **salón** n ‘room, hall’ como si llegara a un salón de clases 1193
1501 **perfectamente** *adv* ‘perfectly’ *sabiendo perfectamente lo que andan diciendo* 1199 s	1510 **cifra** *n* ‘figure, number’ *las grandes cifras macroeconómicas de nuestro país* 1192
1502 **zapato** *n* ‘shoe’ *se limpió el polvo de los zapatos* 1199	1511 **hueso** *n* ‘bone’ *era una osteomielitis del hueso frontal* 1190
1503 **manejar** *v* ‘to handle, manage, drive’ *aprendí a manejar aquél coche* 1197 s	1512 **monte** *n* ‘mountain’ *él atravesó aquellos montes y llanuras* 1189
1504 **brillante** *adj* ‘brilliant’ *ha sido uno de los alumnos más brillantes* 1196 w	1513 **tribunal** *n* ‘court’ *en la jurisprudencia del Tribunal Supremo* 1188
1505 **procedimiento** *n* ‘procedure’ *no aguantaba los procedimientos judiciales* 1195 w	1514 **desconocer** ‘to be unaware of’ *se desconoce qué es lo que hay en el cajón* 1187
1506 **rama** *n* ‘branch’ *saltaba desde la rama de un árbol* 1195	1515 **mensaje** *n* ‘message’ *dejó mensaje en la contestadora* 1186

1507 **comprobar** *v* ‘to prove, check’ *comprobó que su pistola estaba sin seguro* 1195 w	1516 **moneda** *n* ‘coin, currency’ *la propia moneda se convierte en capital* 1185
1508 **contribuir** *v* ‘to contribute’ *una habilidad que contribuye mucho al regocijo* 1194	1517 **relato** *n* ‘story, report’ *el narrador lo dirige a través del relato* 1184

There will also be another index containing the words in alphabetical order, along with part of speech, English gloss, and a cross-reference to the position of the entry in the rank frequency listing (e.g. word number 1081 for *labio*).

Table 13. Alphabetical index in the dictionary

labio *n* lip 1081	**lástima** *n* pity, shame 2574	**legal** *adj* legal 1498
labor *n* work 1404	**lateral** *adj* side, lateral 3723	**legislativo** *adj* legislative 2801
laboratorio *n* laboratory 1751	**latín** *n/adj* latin 2915	**lejano** *adj* distant, far-off 1533
lado *n* side 221	**lavar** *v* to wash 2527	**lejos** *adv* far (away) 624
lago *n* lake 1715	**lazo** *n* tie, bond 3412	**lengua** *n* tongue, language 486
lágrima *n* tear(drop) 954	**le** *pron* to him/her (IO) 27	**lenguaje** *n* language 1125
laguna *n* lagoon, gap, lapse 3155	**leal** *adj* loyal 4602	**lentamente** *adv* slowly 2045
lamentable *adj* regrettable 4191	**lealtad** *n* loyalty 4325	**lente** *n* lens 4641
lamentar *v* to regret 1438	**lección** *n* lesson 3026	**lento** *adj* slow 1539
lamer *v* to lick 5954	**leche** *n* milk 706	**leña** *n* (fire)wood 4670
lámpara 3887	**lecho** *n* (river) bed 2596	**león** *n* lion 1624
lana *n* wool 3551	**lector** *n* reader 1756	**letra** *n* letter 974
lanzar *v* to throw, launch 1229	**lectura** *n* reading (material) 1449	**levantar** *v* to raise, lift 408
lápiz *n* pencil 4829		**leve** *adj* slight, light 2890
largo *adj* long 185	**leer** *v* to read 394	**ley** *n* law 121

Finally, there will be a word class index, with a cross-reference to the entry number for the word in the main frequency listing.

Table 14. Word class index in the dictionary

Adjective		Noun		Verb	
.....		.....		.....	
3659	apasionado	3626	interrupción	3312	ahogar
3662	delicioso	3630	evaluación	3313	fingir
3679	cerebral	3631	serpiente	3317	registrar
3684	templado	3632	capricho	3324	suspender
3686	marítimo	3633	desaparición	3343	dictar
3695	repentino	3634	furia	3349	anticipar
3714	decorativo	3636	revisión	3355	burlar
3719	ardiente	3641	adorno	3362	expulsar
		3644	carreta	3363	distribuir
		3645	cohete	3388	resaltar
Adverb.		**Preposition**			
.....		.....			
3857	ligeramente	15	por		
3881	suavemente	17	con		
3930	francamente	21	al		
4007	enteramente	49	sin		
4071	continuamente	53	sobre		
4162	puramente				

9. The role of technology in determining frequency

The creation of the actual lists for the frequency dictionary is greatly facilitated by advances in technology, which were not available to earlier researchers such as Juilland and Chang-Rodríguez (1964). We have already discussed a number of these factors. For example, a 20 million word corpus in electronic form is quite common today, but would have been completely impossible in the early 1980s, much less in the 1960s. Second, even though it took a few weeks to tag the entire corpus, this is still much less time than five or ten years ago, much less the early 1960s.

Once the tagged output is available, it is then imported into a SQL Server database, where the final frequency data is computed to produce the list of the 6000 most frequent words. This 6000-word list can then be sorted, edited, and manipulated in a number of ways, to produce the different types of indices shown in Tables 12-14 above (frequency, alphabetical, and part of speech). In addition, these databases are made available to research colleagues via the web, so that they can help to enter English glosses and sample sentences from the 20 million word corpus. This information in the database can then be made available to a subsequent group of students via the web, who can correct and modify the original entries. Once the time comes to produce the final frequency dictionary, it is simply a matter of exporting the raw data from the databases into different templates, to create the different frequency indices. Finally, we should note that the modular nature of the data (database tables for tagged output, frequency listings, English glosses and sample sentence, as well as the format templates) means that the data can be modified in one module, without having to explicitly change the other modules.

10. Conclusion

Hopefully the preceding discussion provides some useful insight into the issue of vocabulary range and text coverage, and the way in which the extracted data can be used to create a more useful frequency dictionary of Spanish. From the point of view of a language learner, the important point is that text coverage clearly obeys the law of diminishing returns. With about 4000 words, a language learner would be able to recognize more than 90% of the words in a typical native speaker conversation. If s/he learns two thousand more words, however, this will increase coverage by only about 3-4%. We have also seen that the degree of coverage is a function of register and part of speech, and have provided detailed data to support this view. We have also considered the role of vocabulary range, and how factors such as register affect this as well. Finally, we have briefly considered how this methodology can be interfaced with technology to produce the final output -- an accurate frequency listing of words. Hopefully, this information will be of use not just to linguists and natural language researchers, but to teachers and students alike, who are looking for the most productive way to enhance the acquisition of Spanish vocabulary.

References

Alameda, Juan Ramón and Fernando Cuetos. (1995). Diccionario de frecuencias de las unidades lingüísticas del castellano, Oviedo: Universidad de Oviedo

Buchanan, Milton. (1927). A Graded Spanish Word Book. Toronto: Univ. of Toronto Press.

Davies, Mark. (2002) "Un corpus anotado de 100.000.000 palabras del español histórico y moderno". SEPLN 2002 (Sociedad Española para el Procesamiento del Lenguaje Natural). (Valladolid). 21-27.

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

García Hoz, Victor. (1953). Vocabulario usual, vocabulario común y vocabulario fundamental. Madrid: CSIC.

Garside, Roger, Geoffery Leech and Anthony McEnery. (1997) *Corpus Annotation: Linguistics Information from Computer Text Corpora.* London: Longman.

Jones, Randall. (2003) "An analysis of lexical text coverage in contemporary German". Presentation given at the Corpus Linguistics 2003 conference, Lancaster University (England), March 2003.

Juilland, Alphonse and Eugenio Chang-Rodríguez. (1964). Frequency Dictionary of Spanish Words. The Hague: Mouton.

Nation, Paul. (2001) Learning Vocabulary in Another Language. Cambridge, Cambridge University Press.

Rodriguez Bou, Ismael. (1952) Recuento de vocabulario español. Rió Piedras : Universidad de Puerto Rico.

Sebastián, Nuria, María Antonia Martí, Manuel Carreiras and Fermando Cuetos. (2000), LEXESP, Léxico Informatizado del Español. Barcelona: Ediciones de la Universitat de Barcelona. (CD-ROM only)

Selected Proceedings of the 7th Hispanic Linguistics Symposium

edited by David Eddington

Cascadilla Proceedings Project Somerville, MA 2005

Copyright information

ISBN 1-57473-403-2 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information