Towards a Comprehensive Survey of Register-based Variation in Spanish Syntax

Mark Davies

Brigham Young University

Abstract

This study is based on recent 20 million word corpus of Modern Spanish (1900-1999), containing equivalent sizes of conversation, fiction, and non-fiction. To date, this is the only large, tagged corpus of Spanish that contains texts from a wide range of registers. Nearly 150 syntactic features were tagged, and the frequency of these features in the 20 different registers was calculated. This data is now freely available to researchers via the web. Researchers can examine the frequency of any of the 150 features across the 20 different registers, or examine which of the 150 features are more common in one register than in another. Hopefully this detailed data can be used by teachers and materials developers to provide students of Spanish with a more realistic and holistic view of register variation than has been possible to this point.

1. Introduction

To date there have been no large-scale investigations of register variation in Spanish syntax. It is true that there have been some articles dealing with register variation with individual grammatical constructions (e.g. Davies 1995, Davies 1997, Torres Cacoullos 1999, Davies 2003). There have also been some reference books that provide a study of a wide range of syntactic phenomena of Modern Spanish, but the attention to register differences is often limited and somewhat ad-hoc (e.g. deBruyne 1995, Bosque and Demonte 1999, Butt and Benjamin 2000). In addition, none of the studies that look at more than one syntactic phenomenon is based on a large corpus of Spanish that is composed of many different types of registers. Part of the reason for this is that until very recently, there were no large publicly-available corpora of Spanish that could be used for such analyses.

The lack of in-depth investigations into register variation with a wide range of syntactic phenomena in Spanish is somewhat disappointing, when one considers the range of materials that are available and the studies that have been carried out in other languages. Taking English as an example, one would find the 1200+ page Longman Grammar of Spoken and Written English (Biber et al 1999), which is based on a 40+ million word corpus of spoken, fiction, newspaper, and academic texts. This grammar is replete with detailed register-based analyses and insightful charts and tables that compare the frequency of

hundreds of syntactic constructions and phenomena in the four different registers. (conversation, fiction, news, and academic writing) The goal, of course, would be to make similar materials available for other languages.

In this paper, we will consider the progress that has been made in compiling data for the first large-scale investigation of register differences in Spanish grammar. This study has been carried out with the support of a grant from the National Science Foundation ((#0214438), and it will eventually result in a large multi-dimensional analysis of register variation in Spanish (similar to Biber 1988). These results from Spanish will allow comparison with multi-dimensional analyses of other languages such as English, Tuvaluan, Somali, and Korean (cf. Biber 1995).

In terms of the outline of this paper, Section 2 briefly introduces the 20+ million word corpus that is the basis for the study. Section 3 discusses the way in which the corpus has been annotated and tagged to extract the needed data. Section 4 considers a freely-available web-based interface that allows users to examine variation for nearly 150 different syntactic features in 20 different registers. Finally, Section 5 discusses some of the more salient and interesting findings from the study, in terms of register-based variation in Spanish syntax.

2. The corpus

The corpus that was used in this study is the largest annotated corpus of Spanish, and the only annotated corpus of Spanish to be composed of texts from spoken, fiction, newspaper, and academic registers. The corpus contains 20 million words of text and comprises the "1900s" portion of the NEH-funded Corpus del Español (www.corpusdelespanol.org), which contains 100 million words of text from the 1200s-1900s (for an overview of this corpus and its architecture, see Davies 2002 and Davies 2003b). The following table provides some details of the composition of the 20 million word corpus used in this study.

Table 1. 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 and 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 de la lengua española contemporánea (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 Corpus lingüístico de referencia de la lengua española en argentina (http://www.lllf.uam.es/~fmarcos/informes/corpus/coarginl.html)
6. From the Corpus lingüístico de referencia de la lengua española en chile (http://www.lllf.uam.es/~fmarcos/informes/corpus/cochile.html)

As can be seen, some care was taken to ensure that the corpus adequately represents a wide range of registers from Modern Spanish. The corpus is divided evenly between spoken (e.g. conversations, press conferences, broadcast transcripts), fiction, and non-fiction (e.g. newspapers, academic texts, and encyclopaedias).

3. Annotating the corpus

3.1 There were essentially three stages in the annotation and tagging of the corpus. The first stage was to identify the register for each of the 4051 texts in the corpus. The list of registers included the following:

SPOKEN: 1. contests 2. debate 3. drama 4. formal conversation 5. formal telephone conversation 6. informal conversation 7. institutional dialogue 8. interviews 9. monologue 10. news 11. sports
WRITTEN: 12. academic texts 13. business letters 14. editorials 15. encyclopedias 16. essays and columns 17. general nonfiction 18. literature 19. general news reportage 20. sports reportage

3.2 The second stage was to identify the syntactic features that we felt might be of interest from a register-based perspective. The following is a partial listing of the nearly 150 features that were tagged and analyzed as part of the study (only a partial listing is given for the final category of [Subordinate Clauses]):

GENERAL: 1. type/token ratio 2. avg. word length
NOUNS: 3. NPs without articles, determiners, or numbers, 4. singular nouns, 5. plural nouns, 6. derived nouns (e.g. -azo, -ión, -miento), 7. proper nouns, 8. Diminutives (-ito), 9. Augmentatives (-isimo)
PRONOUNS: 10. 1st person pronouns, 11. 2nd person tu pronouns, 12. 2nd person ud. pronouns, 13. 1st person pro-drop, 14. 2nd person pro-drop , 15. all 3rd person pronouns except 'se', 16. reflexive se, 17. emoción se, 18. 'se', not passive, reflexive, or "matización" , 19. conmigo/contigo/consigo, 20. lo de, la de, etc., 21. lo + ADJ, 22. all clitics 23. pronominal possessives (e.g., la mía), 24. emphatic possessive pronoun (e.g., hija mía), 25. demonstrative pronouns (e.g., ése)
ADJECTIVES: 26. premodifying adjectives, 27. postmodifying adjectives, 28. predicative adjectives, 29. Color adjectives, 30. Size/quantity/extent adjectives, 31. Time adjectives, 32. Evaluative adjectives, 33. Classificational adjectives, 34. Topical adjectives, 35. quantifiers (e.g., muchos, varias, cada)
OTHER NOUN PHRASE ELEMENTS: 36. definite articles, 37. indefinite articles, 38. premodifying possessives, 39. premodifying demonstratives (e.g., ese)
ADVERBS: 40. Adverbs--Place , 41. Adverbs--Time, 42. Adverbs--Manner 43. Adverbs--Stance , 44. Other -mente adverbs, 45. Other adverbs--not -mente
OTHER NON-VERBAL PARTS OF SPEECH: 46. single-word prepositions, 47. multi-word prepositions , , 48. general single-word conjunctions, 49. other single-word conjunctions, 50. multi-word conjunctions, 51. Causal subordinating conjunctions (e.g. puesto que, ya que), 52. Concessive subordinating conjunctions (e.g. aunque, a pesar de que), 53. Conjunctions of condition and exception (e.g. si, con tal que), 54. exclamations (any exclamation mark)
VERBS: 55. Indicative, 56. Subjunctive, 57. Conditional, 58. Present, 59. Imperfect, 60. Future, 61. Past, 62. Progressive, 63. Perfect, 64. Aspectual verbs, 65. Existential 'haber' , 66. ir a, 67. Verbs of mental perception, 68.

Verbs of desire, 69. Verbs of communication, 70. Verbs of facilitation/causation, 71. Verbs of simple occurrence, 72. Verbs of existence/relationship, 73. Verb + infinitive, 74. Haber + que/de, 75. Other obligation verbs: e.g. deber, tener que, 76. Ser passive with 'por', 77. Agentless ser passive, 78. Se passive with 'por' , 79. Agentless se passive , 80. All main verb 'ser' , 81. All main verb 'estar', 82. Infinitives without preceding verb or article, 83. infinitives as nouns, 84. 'ser' + ADJ + 'que' + SUBJUNCTIVE , 85. 'ser' + ADJ + 'que' + INDICATIVE, 86. 'ser' + ADJ + INFINITIVE , 87. modal + present participle

SUBORDINATE CLAUSES: 88. Sentence initial el que, etc., 89. non-sentence initial el que, etc., 90. relative pronoun que, 91. verb complement que, 92. noun complement que, 93. adjective complement que, 94. comparative que, 95. temporal que, 96. Que clefts with indicative ... 141. Donde relatives w/ conditional, 142. Que verb complements with conditional, 143. CU verb complements, 144. CU questions, 145. Yes/No questions, 146. tag questions

3.3 The third stage was to actually tag the 20 million words in the 4051 texts for each of these 150 parts of speech. This was of course the most time-consuming part of the project. The first step was to create a 500,000 word lexicon for Spanish, which was assembled from various sources. The second step was to carry out a traditional linear scan and tagging of the entire corpus. The general schema that we used to design the tagger was the same as that used to create the English tagger that Biber used to tag the 40 million word Longman corpus (see Biber et al 1999). The tagger relied on a sliding ten word window of text with both left and right checking to resolve ambiguity, and it was a hybrid between a strictly rule-based system and a probabilistically-based tagger. During a period of several months, the automatic tagging was revised manually and corrections were made to the tagger. Although we did not carry out exhaustive calculations of the accuracy of the tagger, the manual revision of several 500 word excerpts in the final stages of tagging suggested that the tagger achieved between 98% and 99% accuracy.

The following selection shows a short sample of what the tagged output looked like. Each of the 20 million lines of text contains 1) the word form 2) part of speech (primary and secondary; e.g. imperfect verb / 3pl) 3) miscellaneous features 4) feature tag (e.g. 'que complement' or 'multi-word preposition') and 5) lemma:

(1)
y ^con+coor+++++_gensingcon_+y+
me ^p1cs+per+++++_1pro_+yo+
enfrenté ^vm+is+1s++++_1prod_indicat_preter_+enfrentar+
otra ^d3fs+ind++++!!+_quant_+otro+
vez ^nfs+com+++++_singn_+vez+
con ^en++++++_1wrdprep_+con+

ella ^p3fs+per+++++_3pro_+ella+
y ^con+coor+++++_gensingcon_+y+
con ^en++++++_1wrdprep_+con+
su ^d3cs+pos+++++_prepos_+su+
vela ^nfs+com++++!!+_singn_+vela+
encendida ^jfs+++++!!+_postadj_+encendido+

After the traditional linear tagging, we imported the data into a relational databases (MS SQL Server) where additional disambiguation was carried out. Again, this disambiguation was both rule and probability-based. An example of the probabilistic tagging was the way in which we handled Noun+Past Participle strings, where it is unclear whether the past participle is an adjective (*niños cansados* "tired children", *ventanas rotas* "broken bottles") or the verb in a passive sense (*libros publicados en 1974* "books published in 1974", *dinero gastado ayer* "money spent yesterday"). Using the relational database, we calculated the relative frequency with which each past participle form was used with *ser* "to be" (implying the norm) or *estar* "to be" (implying change from the norm). Typically, past participles occurring more with *estar* lent themselves more to an adjectival interpretation in N+PP sequences, whereas those that occurred more with *ser* lent themselves more to a passive interpretation. In this case, then, the data from one table (relative frequency of PP + *ser/estar*) was used to probabilistically tag sequences in another table (N + PP). Many such updates and corrections to the corpus were made over a period of three months.

4. Web-based interface to register-based differences in syntax

Once the 20 million words in the 4000+ text files were tagged, we then created statistics to show the relative frequency of the 150 features in each of the 20 registers. This data was then imported into a MS SQL Server database, where it was connected to the web. The interface that was created as a result of this processed (now located at http://www.corpusdelespanol.org/ registers/) allows for a wide range of queries by end-users.

4.1 The first basic type of query is to see the relative frequency of one of the 150 syntactic features in each of the 20 registers. Using a drop-down list, users select one of the 150 features and they then see a table like the following (note that all figures for the following four tables have been normalized for frequency per thousand words):

Table 2. Register differences for [first person pronouns]

REGISTER	PER 1000	TOKENS	# WORDS IN REG
SP-informal conversation	19.41	12828	660750
SP-drama	18.76	9419	502044
SP-contests	16.97	1100	64817

REGISTER	PER 1000	TOKENS	# WORDS IN REG
SP-formal conversation	16.77	49363	2942861
SP-debate	14.73	1640	111328
SP-formal telephone conversation	11.25	98	8708
WR-literature	10.10	92998	9210325
SP-interviews	9.42	14551	1544067
SP-institutional dialogue	7.63	4026	527345
WR-business letters	7.62	335	43979
SP-monologue	7.28	2919	401145
SP-news	6.17	516	83664
WR-editorials	4.78	394	82511
SP-sports	4.56	273	59857
WR-essays and columns	3.62	7941	2192407
WR-news reportage	2.28	4767	2094657
WR-general nonfiction	1.57	3608	2293820
WR-academic texts	0.72	146	202943
WR-encyclopedias	0.08	231	2852860

The table shows the actual number of tokens in each register, as well as the normalized value (per thousand words) in each of the 150 registers, and then sorts the results in descending order of frequency.

As the preceding table shows, the use of first person pronouns is the most common in informal conversation and drama and least common in academic texts and encyclopaedias (which is probably not too surprising). Often the findings are less intuitive, as in the following table, which shows the relative frequency of conditional verbs:

Table 3. Register differences for [conditional verbs]

REGISTER	PER 1000	TOKENS	# WORDS IN REG
SP-formal telephone conversation	2.30	20	8708
SP-interviews	2.20	3399	1544067
SP-debate	2.12	236	111328
SP-drama	2.01	1010	502044
SP-monologue	1.90	764	401145
WR-literature	1.85	17004	9210325
SP-institutional dialogue	1.80	947	527345
WR-essays and columns	1.74	3819	2192407
SP-formal conversation	1.70	4994	2942861
WR-news reportage	1.69	3535	2094657
WR-editorials	1.55	128	82511
SP-contests	1.47	95	64817
WR-general nonfiction	1.45	3327	2293820
SP-news	1.35	113	83664
SP-informal conversation	0.97	642	660750
SP-sports	0.97	58	59857
WR-academic texts	0.80	162	202943
WR-encyclopedias	0.63	1805	2852860
WR-business letters	0.00	0	43979

As this table shows, the use of the conditional verb tense tends to be more common in the spoken registers than in the written registers, although there are some spoken registers where it is not very common (e.g. sports broadcasts and informal conversation) and some written registers where it is relatively more common (fiction and essays).

4.2 The website offers an alternative way of searching the data as well. Users can select any two of the twenty registers, and then see which of the 150 syntactic features are used more in Register 1 than in Register 2. For example, the following table shows the listing that compares academic texts to formal conversation. The table shows the frequency (per thousand words) in the two competing registers, and the difference between the two. For example, the first line of the chart indicates that postnominal past participles (*los libros escritos* "the (written) books (written)") occur more than eleven times as frequently in the academic register than in conversation.

Table 4. Syntactic features: [ACADEMIC] vs. [FORMAL CONVERSATION]

FEATURE	DIFF	ACAD	CONV
postnominal past participles	11.17	2.14	0.18
ser passive with 'por'	5.73	0.45	0.07
agentless *ser* passive	4.74	1.70	0.35
topical adjectives	3.08	5.20	1.68
derived nouns (e.g. –azo, -ión, -miento)	3.02	53.22	17.62
postmodifying adjectives	2.87	39.24	13.65
se passive with 'por'	2.83	0.29	0.09
premodifying adjectives	2.38	11.47	4.81
time adjectives	2.29	3.68	1.60
consigo	2.27	0.04	0.01
ser + ADJ + INFINITIVE	2.18	0.30	0.13
infinitives as nouns	2.16	0.64	0.29
agentless se passive	2.16	4.68	2.15
NPs without articles, determiners, or numbers	1.97	101.02	51.39

As this table indicates, [ACADEMIC] texts have (in relative terms) many more passives, nouns, adjectives, and prepositions than [FORMAL CONVERSATION], due to the more "informational" nature of academic texts vis a vis the "interactive" nature of conversation (cf. Biber 1993).

Conversely, one would find the following features to be more common in conversation than in the academic register. Note that many of these features reflect a more "interactive", "people-oriented" type of speech (note also that when the academic figure is .00, it has been smoothed to .01 to avoid division by zero)

Table 5. Syntactic features: [FORMAL CONVERSATION] vs [ACADEMIC]

FEATURE	DIFF	CONV	ACAD
tag questions	295.02	2.95	0.00
2nd person ud. pronouns	143.57	1.44	0.00
exclamations (any exclamation mark)	90.72	1.80	0.01
2nd person tu pronouns	49.86	4.18	0.07
diminutives (-ito)	30.45	0.90	0.02
augmentatives (-isimo)	28.26	0.56	0.01
ir a	23.09	2.39	0.09
1st person pronouns	23.00	16.77	0.72
emphatic possessive pronoun (e.g., hija mía)	19.37	0.19	0.00
yes/no questions	9.74	4.99	0.50
progressive	9.03	1.60	0.17
existential 'haber'	8.30	3.85	0.45
adverbs – Place	8.09	4.35	0.53
CU questions	6.77	0.23	0.02
conmigo	6.59	0.07	0.00
1st person pro-drop	5.39	12.13	2.24

One would probably expect to see clear-cut differences in syntactic features between dissimilar registers such as conversation and academic texts. It is interesting, though, to compare more similar types of speech or writing, and still see what syntactic features differentiate the two registers. For example, one might expect [newspaper editorials] to be almost identical with [newspaper essays and columns], but in fact there are subtle differences. The following table shows some of the syntactic features that are more common in editorials than in essays:

Table 6. Syntactic features: [editorials] vs. [essays and columns]

FEATURE	DIFF	EDIT	ESSAY
emphatic possessive pronoun (e.g., hija mía)	2.49	0.16	0.05
pronominal possessives (e.g., la mía)	2.47	0.21	0.07
Augmentatives (-isimo)	2.14	0.51	0.23
Existential 'haber'	2.13	2.52	1.17
temporal que	2.07	0.08	0.03
Other el que with subjunctive	1.78	0.46	0.25
Other el que with indicative	1.76	5.49	3.11
Verbs of desire	1.73	2.64	1.51
Que clefts with indicative	1.72	0.11	0.05
Causal subordinating conjunctions (e.g. porque, ya que)	1.66	2.97	1.78
non-sentence initial el que, etc.	1.64	3.03	1.84
Que headless & sentence relative clauses INDIC	1.55	0.13	0.08

As we see, because of the persuasive nature of editorials we find more emphatic constructions, verbs of desire, and (perhaps due to the need to build up complex series of argumentation) more clefting types of constructions. In summary, because there are 20 different registers in the corpus and because users can

compare any two registers in the list, this allows for nearly 400 different pair-wise comparisons of registers in Spanish.

Finally, in addition to being able to see the frequency of 150 different features in all 20 registers, as well as being able to compare two registers directly, the website also allows users to see a KWIC (keyword in context) display for any of these data. For example, if users want to see examples of the [verbs of desire] that are more common in editorials than in essays (the query just discussed), they simply click on the [verbs of desire] entry in the listing, and they then see KWIC display for the first fifty occurrences in that register (in this case editorials), as in the following:

1. del asesinato de estas palabras. Quiero ser presidente , pero no a
2. cinismo fácil y divertido . No quiero decir que lo sea , cínico
3. vez valga la comparación , pero prefiero otros recuerdos personales . Va para
4. del grupo . Cuántos Sharnu desearíamos ? Cuántos son ? Leo las
5. a obra es muy valiosa y necesitábamos tenerla . Mi juicio es a
6. Amaba y odiaba su obra . espero arruinar el apetito de cada hijo
7. carta a su hijo , pero prefiero escribir de Ana y para Ana
8. impide que veamos lo que no queremos ver , y nos vamos corriendo

To summarize, this is the first and only corpus interface that allows researchers of Spanish to directly examine register differences in Spanish. Because the data is freely available to all researchers, this data will hopefully be used by many people to create more detailed descriptions of Spanish, which can then be used to develop more useful materials for the classroom.

5. Examples of register variation in Spanish

In this section, we will briefly provide two examples of ways in which a cluster of features are distributed differently in competing registers of Spanish. In order to simplify the presentation, we have grouped the 20 individual registers into three "macro" registers – conversation, fiction, and non-fiction.

The first table shows the relative frequency of different parts of speech in these three registers.

Table 7. Relative frequency of different parts of speech

	Percent		
	Spoken	Fiction	Non-fiction
noun	19.5	24.7	32.4
verb	19.4	18.6	12.0
adjective	4.0	4.5	7.2
adverb	10.5	5.8	3.1
pronoun	9.3	7.2	3.1
conjunction	7.0	6.1	5.0
determiner	3.5	3.5	2.7

preposition	12.1	15.0	18.4
article	9.0	11.5	13.9
question word	3.5	2.7	1.6

This table shows, for example, that there are roughly as many nouns as verbs in spoken Spanish (about 19.5 percent of all tokens for each of these two parts of speech). In non-fiction texts, however, there are much more nouns than verbs – almost three times as many. Not surprisingly, the "noun-heavy" non-fiction texts also have more adjectives and more prepositions, while the "verb-heavy" spoken register has more adverbs. This difference is a result of the general "information-oriented" nature of non-fiction texts, compared to the "interactive nature" of conversation (cf. Biber 1993). Note also that the fiction texts in general occupy a position between conversation and non-fiction. Finally, we note that these data tend to agree quite well with the relative frequency of different parts of speech in English (for example, cf. Biber et al. 1999: 65-69).

The second example of register variation deals with the relative frequency of the different verb tenses in each of the three macro registers, and the data for these features are found in the following table:

Table 8. Relative frequency of different verb tenses

	Percent		
	Spoken	Fiction	Non-fiction
Indicative			
present	61.3	33.6	45.8
preterit	11.0	23.8	30.2
imperfect	13.6	26.8	13.4
future	0.8	1.5	0.7
conditional	1.4	1.9	1.0
perfect	3.9	1.4	3.1
pluperfect	0.7	2.8	1.4
Subjunctive	5.8	7.4	4.3
Present	4.2	3.3	2.9
Imperfect	1.3	3.6	1.3
Perfect	0.1	0.1	0.1
Pluperfect	0.2	0.6	0.1
Progressive	1.4	0.7	0.2

These data provide a number of insights into register variation in Spanish. First, they show that the two primary past tenses (preterit and imperfect) account for more than 50% of all verbs in fiction, which is more frequent than non-fiction texts and more than twice as common as conversation. This compares nicely with the data for English (found in Biber 1993), who explains that fiction texts of course contain more past tense verbs because they are more oriented towards

narrated past events, whereas conversation is oriented more towards the present. Finally, this basic distinction between the present and the past also carries over into compound verb tenses, such as the perfect (present-oriented) and the pluperfect (past-oriented).

The second major difference deals with aspect – specifically the relative frequency of the progressive. As the table indicates, the progressive is most frequent in spoken Spanish, followed by fiction, and finally by non-fiction, where it has only about one-seventh the frequency of spoken texts. Following Biber et al (1999: 461-62) this is due to the "ongoing, here-and-now" nature of conversation, as opposed to non-fiction texts, which tend to deal more with general relationships outside of any particular temporal frame.

The third major difference deals with mood in Spanish, which of course is much more marked in Spanish (via the subjunctive) than it is in English. As the table indicates, the subjunctive mood is the most common in fiction, then spoken, and then non-fiction. This distinction is perhaps somewhat less intuitive than the preceding two features. The higher frequency of the subjunctive in fiction may be due to the need to explicitly spell out feelings and desires and opinions of the protagonists in the story (and these types of verbs are the primary triggers for the subjunctive in Spanish), vis a vis conversation, where these are implied as part of the speech act. Finally, the higher frequency of the subjunctive in fiction and conversation than in non-fiction texts may be due to the "people-oriented" nature of the first two texts, where the attitudes and feelings of one person affect a second person, which is a major motivation for the subjunctive (cf. Butt and Benjamin 246-56).

6. Conclusion

While other languages such as English have detailed studies of register differences (e.g. Biber et al 1999), such insights have not been readily available for Spanish. To this point, students, teachers, and materials developers for Spanish have had to simply rely on intuition to understand how spoken Spanish differs from written texts, and how the different registers (formal and informal conversation, fiction, academic texts) relate to each other. With the data from the present study, however, researchers and students of Spanish will finally have access to a wealth of information – via a free and simple web-based interface – which will provide them with a much-improved understanding of the precise nature of syntactic variation in Spanish.

References

Biber, D. (1988), *Variation across speech and writing.* Cambridge: Cambridge University Press.

----------. (1993), 'The multi-dimensional approach to linguistic analyses of genre variation: an overview of methodology and findings', *Computers and the Humanities* 26: 331-45.

----------. (1995), D*imensions of register variation: A cross-linguistic comparison.* Cambridge: Cambridge University Press.

----------, S. Johansson, G. Leech, S. Conrad, E. Finegan. (1999), *The Longman grammar of spoken and written English.* London: Longman.

de Bruyne, J. (1995), *A Comprehensive Spanish Grammar.* Oxford: Blackwell.

Bosque, I. and V. Demonte. (1999), *Gramática descriptiva de la lengua española.* 3 vols. Madrid: Espasa Calpe.

Butt, J. and C. Benjamin. (2000), *A New Reference Grammar of Modern Spanish.* New York: McGraw-Hill.

Davies, M. (1995), 'Analyzing Syntactic Variation with Computer-Based Corpora: The Case of Modern Spanish Clitic Climbing', *Hispania* 78:370-380.

----------. (1997) 'A Corpus-Based Analysis of Subject Raising in Modern Spanish', *Hispanic Linguistics* 9: 33-63.

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

----------. (2003a), 'Diachronic Shifts and Register Variation with the "Lexical Subject of Infinitive" Construction. (Para yo hacerlo)', in: S. Montrul and F. Ordóñez (eds.) *Linguistic Theory and Language Development in Hispanic Languages.* Somerville, MA: Cascadilla Press. 13-29.

----------. (2003b), 'Relational n-gram databases as a basis for unlimited annotation on very large corpora', in: K. Simov (ed.) *Proceedings from the Workshop on Shallow Processing of Large Corpora* (Lancaster, England, March 2003). 23-33.

Torres Cacoullos, Rena. (1999), 'Construction frequency and reductive change: diachronic and register variation in Spanish clitic climbing', *Language Variation and Change* 11:143-170.