Technological Applications to Linguistic Research: A Corpus Analysis of Frequency Adverbials

Roberto Mayoral Hernández and Asier Alcázar
University of Alabama-Birmingham and University of Missouri-Columbia

1. Introduction[1]

Our investigation follows modern trends that incorporate technological advances to the study of language by means of corpora (Marcos Marín, 1995; Biber, Davies, Jones, & Tracy-Ventura, 2006; Baayen, in press; Alcázar, 2008). We extracted the data from *Corpus de Referencia del Español Actual* (CREA), the largest corpus of Spanish available online. In his opening plenary talk for HLS 2007, José Moreno de Alba stressed the importance of text digitalization to create free online corpora, which have become an invaluable resource for linguistic research. The paper also benefits from an interdisciplinary approach that combines traditional sociolinguistic analysis with computational linguistics (Gries, 2003; Levy, 2008). In this line, we developed computational tools to automatically annotate relevant information contained in CREA. Finally, our study employs statistical methods to test the significance of linguistic hypotheses (Tagliamonte, 2006).

This article looks at the position of adverbial expressions in Spanish. The term "adverbial expression" refers to a very heterogeneous group of grammatical categories, like adverbs, prepositional phrases and nominal phrases. This group is also characterized by various syntactic and semantic properties (see Kovacci (1999) for a detailed discussion). In this paper, we focus on the ordering of Spanish frequency adverbials, following previous research by Mayoral Hernández (2004, 2008). Specifically, his manually annotated corpus has been extended by means of computational tools to include sociolinguistic factors.

Although previous analyses propose that a difference in position correlates with a difference in meaning (e.g. Cinque (1999) and other generative approaches), we will contend that a purely semantic or syntactic analysis yields unsatisfactory predictions. In fact, we defend the idea that the distribution of adverbials is sensitive to speaker preferences and universal constraints, as proposed in the psycholinguistic literature (Wasow & Arnold, 2003; Lohse et al., 2004). Here we will adopt a variationist approach in which the interaction of different factors determines the position of adverbials.

The variationist approach assumes that variation arises from a combination of different factors. These factors, or independent variables, are often of a diverse nature. Thus, it is desirable to combine in the same model variables that pertain to various levels of representation: phonological, semantic, pragmatic, syntactic, etc. As a result, sometimes it becomes irrelevant to try and elucidate if a specific grammatical phenomenon should be better described as syntactic, morphological, phonetic, semantic, etc; but rather, it emphasizes the necessity of describing every language process as a result of the interaction of cognitive, mental and purely grammatical constraints (Hawkins, 1994, 1999, 2000, 2001).

In this paper, we will examine how three sociolinguistic and stylistic factors can influence the ordering of adverbials, in particular *gender*, *language variety* and *genre*. Abundant research has proven that these three variables play a center role in many grammatical phenomena (e.g., Silva-Corvalán, 2001; Lavandera, 1975; Coulmas, 2001), as we will see in the following sections.

The outline of the paper is as follows. Section 2 describes the adverbial alternation in Spanish. Section 3 introduces a quantitative analysis of three factors that affect adverbial ordering: gender,

[1] Thanks to the audiences at HLS 2006, LSRL 36 and HILSA workshops for useful comments and suggestions. We are especially indebted to Carmen Silva-Corvalán for extensive comments on an earlier version of this paper. We also thank Lotfi Sahayi and Scott Schwenter for helpful discussions. All errors are our own.

language variety and genre. Section 4 details the sequence of steps followed to extend the manually annotated corpus using computational tools. Section 5 provides the results of our experiment. Section 6 discusses the implications of the analysis for linguistic research. Finally, section 7 presents the conclusions of our paper.

2. Adverbial alternation

In Spanish, frequency adverbials like *frecuentemente* "frequently" can appear in either preverbal or postverbal position. The following examples are actual written samples of Spanish that we extracted from the online corpus CREA. Each sentence illustrates one of the four positions considered in our study.

(1) Preverbal 1: Before [co-occurring XP]
Frecuentemente [*los miembros de las comunidades*] *reciben cursos de protección ambiental*
"Frequently the members of the communities receive courses on environmental protection"

(2) Preverbal 2: after [co-occurring XP]
[*Los agentes del SIN*] ***frecuentemente*** *realizan redadas en empresas…*
"The SIN agents frequently carry out raids on companies…"

(3) Postverbal 1: before [co-occurring XP(s)
la actividad del citado empresario transciende ***frecuentemente*** [*el mero aspecto comercial*]
"the activity of the aforementioned businessman frequently transcends the merely commercial aspect"

(4) Postverbal 2: after [co-occurring XP(s)]
La situación ha sido [*muy tensa*] ***frecuentemente***
"The situation has frequently been very tense"

In these sentences, the frequency adverbial could have been expressed in any of the other positions without an apparent change in meaning.

However, the literature dealing with Spanish adverbs mentions that preverbal and postverbal adverbials have different interpretations that can be also observed in a different syntactic behavior. In her exhaustive description of Spanish adverbials, Kovacci (1999) assigns different properties to preverbal and postverbal positions. In particular, she explains that postverbal adverbials (i) modify the verb, having a circumstantial use, (ii) imply the text without the adverb and (iii) can be paraphrased using a sentence introduced by *como* "how" or *cuando* "when". By way of example, a sentence like (5a), containing the postverbal adverb *frecuentemente* "frequently", should entail the same sentence with no adverb, as in (5b), and it should be possible to paraphrase it by using a sentence headed by the complementizers *como* "how" or *cuando* "when", as in (5c).

(5) a. *mis amigos comen patatas frecuentemente*
"my friends frequently eat potatoes"
b. *mis amigos comen patatas*
"my friends eat potatoes"
c. *es frecuentemente cuando/como mis amigos comen patatas*
"It's frequently when my friends eat potatoes"

Kovacci assumes that frequency adverbials show different properties when they occur in preverbal position: (i) they modify the sentence, (ii) do not imply the sentence without the adverb and (iii) cannot be paraphrased by a clause with *cuando* "when" or *como* "how". But this does not seem to be the case. A sentence with a preverbal adverbial like (6a) also entails the same sentence without the adverb, as in (6b), and also admits paraphrases with *cuando* "when" or *como* "how", as in (6c). The grammaticality of sentences (6b) and (6c) seems to indicate that the difference between preverbal and postverbal

frequency adverbials, if any, might not be derived exclusively from purely semantic or syntactic properties.

(6) a. *frecuentemente mis amigos comen patatas*
"my friends frequently eat potatoes"
b. *mis amigos comen patatas*
"my friends eat potatoes"
c. *es frecuentemente cuando/como mis amigos comen patatas*
"It's frequently when my friends eat potatoes"

The previous examples show that a change in the position of the adverb does not necessarily entail a change in meaning. Even if there were preferred interpretations associated with different positions, psycholinguistic research has shown that the avoidance of meaning ambiguities is not a factor that triggers ordering alternations in language production (Hawkins, 2000; Wasow & Arnold, 2003). A purely syntactic or semantic theory cannot account for the wide variation in grammaticality judgments existing with respect to preferred orders. Nevertheless, we do not intend to hold that there cannot be meaning differences associated with different positions. We acknowledge that in certain contexts, like the ones studied by Kovacci containing quantifiers, it is possible to find semantic differences between preverbal and postverbal adverbials, or at least a predisposition to interpret adverbial scope in the way Kovacci indicates. Here we will defend that adverbial position is related to stylistic preferences and sociolinguistic forces.

Previous research has shown that several factors influence the position of frequency adverbials. Using a variationist approach, Mayoral Hernández (2004, 2008) showed that "weight", "subject position" and "verb type" can determine the final collocation of constituents in the sentence. Based on Hawkins' research, Mayoral Hernández demonstrated that weight, counted as number of words, is a predictive factor for the collocation of adverbial expressions and other constituents in the sentence: heavier elements, i.e., constituents with a higher number of words, will tend to appear in sentence final position, whether they are adverbials or not. As Hawkins explains, the concept of weight can only be relative, since the different weights of the constituents in the sentence will interact to make lighter elements precede heavier ones.

The presence of overt subjects in the sentence, or their absence, has also interesting effects on the position of adverbials. In sum, frequency adverbials and subjects often occur in complementary distribution, as frequency adverbials tend to gravitate to preverbal positions in sentences that lack an overt subject.

The third factor studied by Mayoral Hernández (2008) is the type of verb. When comparing transitive, copulative, unergative and unaccusative verbs, he found that unaccusative verbs are associated with a higher percentage of preverbal adverbials. This is partially due to a higher percentage of postverbal subjects in unaccusatives.

In the following section we will add three more factors to Mayoral Hernández's previous work.

3. Statistical analysis

Our research expands that of Mayoral Hernández (2004, 2008) by testing whether sociolinguistic and stylistic factors also affect the position of frequency adverbials. Here we address three novel research questions in the domain of the adverbial alternation: (i) do women and men show the same syntactic distribution? (ii) are there dialectal differences between Latin American and Peninsular varieties? (iii) is genre a good predictor of syntactic position?

The three relevant independent variables to analyze are gender, language variety and genre. The rest of this section deals with the three hypotheses associated to each factor.

3.1. First hypothesis

Gender has been studied as a possible source of sociolinguistic variation for a variety of grammatical phenomena (Labov, 1966; for Spanish, see Silva-Corvalán, 2001 and references therein).

For instance, Lavandera (1975) analyzes the use of imperfect indicative and simple conditional tenses in the main clause of conditional sentences in Buenos Aires. Table 1 shows that women and men have different preferences for tense in the apodosis of conditional sentences. Men prefer the conditional form in *-ría*, while women prefer the imperfect form in *–ba*.

Table 1: *Conditional vs. imperfect in apodosis by gender*
(Lavandera, 1975 cf. Silva-Corvalán, 1989)

	N Total	**-ría**		**-ba**	
Total speakers	33	19	55%	14	45%
Women	19	8	35%	11	65%
Men	14	11	79%	3	21%

Considering the gender differences reported in the literature, our first hypothesis states that gender will influence the position of adverbials. According to this hypothesis, women and men will show a different syntactic distribution. Gender is a dichotomous variable with two values: (i) male and (ii) female.

3.2. Second hypothesis

Numerous dialectology studies have shown salient differences in the phonology and syntax of Peninsular and Latin American Spanish (Silva-Corvalán, 1997; Morales, 1986; Ranson, 1991).

Table 2 presents the relevant distribution of subject expression as seen in three varieties of Spanish. Peninsular Spanish, represented by Ranson's (1991) analysis of Andalusian Spanish, shows a higher tendency to omit subjects in third person when compared with American varieties (Los Angeles and Puerto Rico). One might think that this difference is due to phonological reasons. The omission of final /s/ creates meaning ambiguities in the absence of an explicit subject in Latin American varieties. By way of example, the second person singular *comes* 'you eat' is pronounced in the same way as the third person singular *come* '(s)he eats'. Hence, a higher percentage of overt subjects would help reduce this ambiguity. However, it is important to note that the Spanish spoken in Andalusia, Southern Spain, shares the aspiration and omission of the phoneme /s/ in coda position with American varieties. This stands in contrast to Northern and Central varieties in Spain, which retain this phoneme. Overall, table 2 shows that, when varieties with similar processes of consonant reduction are compared, there are still significant differences pertaining to language variety[2].

Table 2: *Subject expression across Spanish varieties*
(Silva-Corvalán, 1997; Morales, 1986; and Ranson, 1991, respectively)

	Los Angeles	**Puerto Rico**	**Andalusia**
Yo	42%	47%	50%
Él/Ella	31%	37%	10%
Nosotros/as	18%	19%	19%
Ellos/as	18%	18%	9%

[2] An anonymous reviewer mentions that the table is incomplete because it lacks data for 2nd person. We took the table from Silva-Corvalán (1997) and it did not include this information. The major point of discussion in this section is to illustrate that variation exists among different varieties of Spanish. The reviewer also suggests that the table should conform to the other tables in the paper and include N and chi-square values. However, we do not have access to the original data reported in Silva-Corvalán (1997).

In view of the differences between Peninsular and Latin American Spanish, the second hypothesis predicts that adverbial position will also be influenced by this factor.

The factor *language variety* is a dichotomous variable with two values: (i) Peninsular and (ii) Latin American. The variable *Peninsular* represents documents from Spain. The variable *Latin American* represents 21 countries: all Latin American countries and the USA.

The original study by Mayoral Hernández (2004) randomly searched the CREA corpus for sentences containing frequency adverbials. Accordingly, the data sample did not result in a homogenous distribution of tokens by country. In effect, texts from Spain are overrepresented (897 n = 56.2%) while Mexico is the Latin American country with most tokens (211 n = 13.2%). This is due to the preponderance of texts from Spain in CREA. Evidently, Latin American Spanish is not a uniform variety. A simple look at the literature dealing with Spanish dialectology will show that neither Latin American nor Peninsular Spanish are homogenous (Moreno Fernández, 1993; Fontanella de Weinberg, 1992; Lipski 1994). Future research will incorporate additional tokens from Latin American varieties to further explore regional variation.

Having said that, the division that we adopt here can be defended on methodological, historical, geographical and grammatical grounds. First of all, we are only dealing with written Spanish as it appears in books and press —a formal and educated style. Written Spanish should not be affected by the multiplicity of phonetic phenomena documented across different varieties of spoken Spanish. Furthermore, even though some degree of lexical divergence is naturally expected, our research does not deal with lexical variation, since very specific adverbs have been selected and controlled.

3.3. Third hypothesis

Genre differences often merit the attention of the specialized literature. For example, Finegan and Biber (2001) analyze different cases in which genre has a direct effect on different grammatical features in English, such as contractions, *that* omission and the use of the pronoun *it*.

As presented in table 3, press reportage and academic prose diverge significantly, particularly in the use of contractions and the omission of the complementizer *that*.

Table 3: *Overview of situational variation (per thousand words)*
(Finegan and Biber, 2001)

		Contractions	***That* omission**	**Pronoun *it***
Written	**Press reportage**	1.8	2.0	5.8
	Academic prose	0.1	0.4	5.9

The third hypothesis predicts that genre will play a role in the ordering of adverbials. In our study we focus on the divergence between writing styles in books and newspapers. Thus, genre is a dichotomous variable with two values: (i) press and (ii) book.

4. Methodology

As mentioned before, our research extends the variationist study on frequency adverbials by Mayoral Hernández (2004, 2008). The original corpus consists of 1033 sentences from the online corpus CREA. Each sentence contains one of three frequency adverbials that were selected to represent different weights (different number of words): (a) *frecuentemente* “frequently”, (b) *en muchas ocasiones* “on many occasions” and (c) *en más de una ocasión* “on more than one occasion”.

This manually annotated corpus contained some information that was not utilized in the original study. This information is part of the search results provided by CREA and includes (i) the publisher, (ii) the year of publication, (iii) the country of origin or language variety, (iv) the document source or

genre (primarily books and press, but also oral and electronic texts in a few cases), (v) the author, and (vi) the topic[3] (agriculture, politics, medicine, economy, etc.).

From the above mentioned fields, we extracted the last four (iii-vi). This information had not been previously annotated, since the focus in the work leading to the creation of the corpus was to determine the relevance of factors such as adverb weight and type of verb. Using a series of computer programs written in Python and specifically designed for this research, we were able to automatically annotate these fields and add them to the original manual annotation.

The information relative to country (iii) provided by CREA was annotated as our independent variable language variety. We grouped together all the Latin American countries in our corpus under the factor Latin American varieties, while Spain was coded as Peninsular varieties.

Regarding the variable genre (iv), a filter excluded instances of oral and electronic texts (web pages, email communications). The remaining genres, namely academic and literary books, magazines and press articles, were merged into two naturally occurring categories: books and press.

Finally, the author variable (v) required manual annotation as male or female for the first occurrence in the corpus. In cases of multiple authorship, we required that all authors belong in the same gender category to call for the label male or female. The mixed cases we coded as mixed and not included in this analysis. Note that the press documents did not contain author information, and we excluded them from the investigation.

It turned out that the original corpus had a much higher number of male authors. To counterbalance this bias, an additional 564 sentences by female authors were added to balance the male-female ratio, increasing the corpus to a total of 1597 sentences. These sentences were also annotated automatically for the variables language variety and genre.

Finally, one of the Python programs converted the enriched annotation into an SPSS readable file.

Given our earlier experiences in manual annotation, we found the aid of Python a powerful tool to speed up the coding process. It allowed us to add new tokens easily when the need arose (male author bias). Finally, these programs can be reused for further variation studies with CREA[4].

5. Results

In this section, we present the results in crosstabulation tables and apply the Pearson's chi-square test to determine statistical significance. We have adopted the standard probability value as the threshold for significance, and therefore any $p < 0.05$ will be considered statistically significant.

5.1. Results: Gender

Table 4 shows that gender is statistically significant ($p = 0.0001$). Women favor postverbal adverbials (62.5%), while men prefer preverbal positions (45.2%). The reader will observe that the total number of tokens in this table is only 657, rather than the 1597 sentences that constitute the entire corpus. This is due to an idiosyncratic characteristic of the author annotation in CREA, which contains this information for books only ($n = 657$), as opposed to magazine and newspaper articles. An anonymous reviewer points out that "it needs to be noted that several editors (of both sexes) of the publishing houses that produce the books and periodicals may actually change the position of the adverb in the editing process." Although we agree with the reviewer that the editing process may effect stylistic changes, it is important to stress that the position of frequency adverbials would be unlikely for an editor to target. This research could be complemented with the study of oral production. However, the present study could not be fully replicated because of fundamental differences between written and spoken language. For example, Mayoral Hernández (2008), also using CREA, finds that

[3] The independent variable Topic proved to be statistically significant as well. However, we need to focus our discussion on the fields (iii-v) for reasons of space.

[4] For an anonymous reviewer section 4 would benefit from an illustration of the online corpus CREA with screenshots as well as a more extensive description of the Python programs. Unfortunately, for reasons of space, we will not be able to extend this section further. For more information, the reader is referred to http://corpus.rae.es/ayuda_c.htm

the presence of an overt subject is a statistically significant factor: while preverbal subjects favor postverbal adverbials, postverbal subjects tend to occur with preverbal adverbials. Since in spoken language subjects are almost systematically omitted, the study could only be replicated with a rather large corpus. On a different note, a diachronic analysis would be prevented in the absence of spoken data from past centuries (Mayoral Hernández & Alcázar, 2008).

Table 4: *Position of adverbials by gender*

			Gender		Total
			Female	Male	
Position of adverbials	Postverbal Position	Count	195	156	351
		% within gender	62.5%	45.2%	53.4%
	Preverbal Position	Count	117	189	306
		% within gender	37.5%	54.8%	46.6%
Total		Count	312	345	657
		% within gender	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	19.667	1	.0001

5.2. Results: Language variety

This independent variable proved to be significant as well (p = 0.0001). As table 5 indicates, Peninsular varieties favor postverbal adverbials while Latin American varieties show no clear preference. All the sentences included in our corpus (n = 1597) contain information relative to what we refer to as language variety, which is indicated as "country" in CREA notation.

Table 5: *Position of adverbials by language variety*

			Language Variety (LV)		Total
			Peninsular	Latin American	
Position of adverbials	Postverbal Position	Count	512	335	847
		% within LV	60.0%	50.9%	56.0%
	Preverbal Position	Count	342	323	665
		% within LV	40.0%	49.1%	44.0%
Total		Count	854	658	1512
		% within LV	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	12.331	1	.0001

5.3. Results: Genre

Genre also turned out to be statistically significant (p = 0.008). Both books and press favor postverbal positions, although table 6 shows that newspaper authors make use of a significantly higher number of postverbal adverbials. The total number of sentences that fit into the book or press category is 1512, which is almost the complete corpus. The rest of sentences (n = 85) are examples of spoken language (i.e., transcriptions of radio and TV programs, court transcripts, etc.).

Table 6: *Position of adverbials by genre*

			Genre		
			Book	Press	Total
Position of adverbials	Postverbal Position	Count	388	459	847
		% within genre	52.6%	59.3%	56.0%
	Preverbal Position	Count	350	315	665
		% within genre	47.4%	40.7%	44.0%
Total		Count	738	774	1512
		% within genre	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	6.940	1	.008

5.4. Results: Gender by Language variety

The results presented thus far may have potential interactions among the different factors. It is important to control for these potential interactions to avoid unreliable results. In this section, we have separated the gender data for the two language varieties that we study in this article: Peninsular and Latin American Spanish.

Table 7 shows gender information for Peninsular Spanish only (n = 343). This table indicates that both men and women prefer postverbal positions: 62.8% and 53.5%, respectively. The difference between the two groups is not statistically significant (p= 0.088). However, men show a strong tendency for preverbal positions. In fact, men have a more balanced distribution between preverbal and postverbal adverbials.

Table 7: *Position of adverbials by gender in Peninsular Spanish*

			Gender		
			Female	Male	Total
Position of adverbials	Postverbal Position	Count	91	106	197
		% within gender	62.8%	53.5%	57.4%
	Preverbal Position	Count	54	92	146
		% within gender	37.2%	46.5%	42.6%
Total		Count	145	198	343
		% within gender	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	2.913	1	.088

The results obtained for Peninsular Spanish contrasts with the overall gender results obtained before, which were significant. At this point, we predict that the overall gender significance obtained for gender is due to the influence of the Latin American data. In effect, table 8 reveals this influence. Table 8 shows gender information for Latin American Spanish only (n = 314). The influence of gender on adverbial ordering is statistically significant in Latin America (p= 0.0001).

Thus, the overall significance obtained in the general gender table results from the higher percentage of preverbal positions used by Latin American men. Although it is important to remember that Spanish men also show a higher percentage of preverbal adverbials when compared to women.

			Gender		
			Female	Male	Total
Position of adverbials	Postverbal Position	Count	104	50	154
		% within gender	62.3%	34.0%	49.0%
	Preverbal Position	Count	63	97	160
		% within gender	37.7%	66.0%	51.0%
Total		Count	167	147	314
		% within gender	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	24.988	1	.0001

Table 8: *Position of adverbials by gender in Latin American Spanish*

5.5. Results: Genre by Language variety

As in the previous section, here we control for potential interactions among factors and provide two different tables for each language variety. Table 9 shows genre information for Peninsular Spanish only (n= 854), where genre is statistically significant (p= 0.018). As seen in this table, both books and press favor postverbal positions in Peninsular Spanish, although it is clear that the style used in press is associated with a more marked preference for this position than it is in books. These results agree with the general genre results obtained in table 6.

Table 9: *Position of adverbials by genre in Spain*

			Genre		
			Book	Press	Total
Position of adverbials	Postverbal Position	Count	233	279	512
		% within genre	55.9%	63.8%	60.0%
	Preverbal Position	Count	184	158	342
		% within genre	44.1%	36.2%	40.0%
Total		Count	417	437	854
		% within genre	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	5.644	1	.018

However, when only Latin American countries are analyzed (n = 658), there seems to be no significant variation associated to these two genres. Table 10 indicates that there is no significant preference for either preverbal or postverbal positions associated to different genres (p= 0.189).

Table 10: *Position of adverbials by genre in Latin America*

			Genre		
			Book	Press	Total
Position of adverbials	Postverbal Position	Count	155	180	335
		% within genre	48.3%	53.4%	50.9%
	Preverbal Position	Count	166	157	323
		% within genre	51.7%	46.6%	49.1%
Total		Count	321	337	658
		% within genre	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	1.728	1	.189

6. General discussion and conclusions

The statistical analysis initially shows that gender is significant overall (p = 0.0001), i.e., when both Peninsular and Latin American varieties are merged together. Nevertheless, once we control the results by language variety, we observe that gender in Latin American data is significant (p= 0.0001), while gender in Peninsular data is not (p = 0.088). These results indicate that the Latin American data drove the overall significance of gender. In fact, women behave alike across varieties, with very similar percentages that range from 62.3% to 62.8%.

Latin American men are the main source of variation because they clearly favor preverbal positions (66.0%). Although Spanish men pattern together with women in that both groups prefer postverbal adverbials (53.5% and 62.8%, respectively), the men show more of a tendency towards preverbal positions (46.5%) than the women.

The differences observed in the domain of adverbial position pose the following question: who is driving the change, men or women? The sociolinguistic literature notes that women tend to be more conservative than men (Labov 1966, Silva-Corvalán 2001). In light of this observation, we could hypothesize that women display the more conservative variant and that it is men who innovate. In a parallel study using the online *Corpus Diacrónico del Español* (CORDE), Mayoral Hernández and Alcázar (2007, 2008) found that the preferred position has been postverbal since the 16th century. It seems reasonable to conclude that women follow the general pattern observed in the sociolinguistic literature. The preference of Latin American men for preverbal positions could run parallel to developments in the history of other languages. It is widely known that languages like French or English have an almost obligatory position for adverbs, which precedes the main verb. Spanish might as well be slowly evolving towards preverbal positions. Nevertheless, testing this hypothesis would require additional research.

After analyzing the previous results, the significance of the variable *language variety* becomes uncertain. We know that the only difference between Peninsular and Latin American Spanish is due to Latin American men. If they were excluded from the analysis, as in table 11, we would be unable to reject the null hypothesis (p = 0.297).

Table 11: *Position of adverbials by language variety, excluding Latin American men*

			Language Variety (LV)		
			Peninsular	Latin American	Total
Position of adverbials	Postverbal Position	Count	197	104	301
		% within LV	57.4%	62.3%	59.0%
	Preverbal Position	Count	146	63	209
		% within LV	42.6%	37.7%	41.0%
Total		Count	343	167	510
		% within LV	100.0%	100.0%	100.0%

	Value	d.f.	Asymp. Sig (2 sided)
Pearson Chi-Square	1.088	1	.297

In any case, the relevance of this factor is by and large contingent upon its interaction with the other two independent variables analyzed, namely *gender* and *genre*, making it difficult to assess its individual weight.

The last factor, genre, proved to be statistically significant overall, when both Peninsular and Latin American data are merged together. But when they are separated, it became relevant only in Peninsular Spanish, but not in Latin American Spanish. This indicates an explicit stylistic difference between the two genres in Spain, which are manifested in different ordering preferences.

To conclude, we have shown that sociolinguistic and stylistic factors (gender, language variety and genre) affect the adverbial alternation. These factors add to those reported by Mayoral Hernández (2004): adverbial weight, presence of other XPs (e.g., subjects), and type of verb. Taken together, the results confirm that the variable position of frequency adverbials cannot be a purely structural

phenomenon (with Mayoral Hernández, 2004, 2008). The syntactic distribution of adverbials is thus not fixed, for it is sensitive to speaker preferences (Wasow & Arnold, 2003; Lohse et al., 2004). Our investigation has taken advantage of new technological applications for the study of language, such as computational tools for text processing and annotation, as well as new linguistic resources like online corpora.

References

Alcázar, A. (2008). Information source in Spanish and Basque: A parallel corpus study. Paper presented at *30th Annual meeting of the German linguistic society* (Workshop on evidentiality in European languages). University of Bamberg, Bamberg, Germany.

Baayen, R. H. (in press). Corpus linguistics in morphology: Morphological productivity. To appear in A. Ludeling, M. Kyto & T. McEnery (Ed.), *Handbook of corpus linguistics* (Handbuecher zur Sprach-und Kommunikationswissenschaft). De Gruyter.

Biber, D., Davies, M., Jones, J. K., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, 1, 1-37.

Coulmas, F. (2001). Sociolinguistics. In M. Aronoff, & J. Rees-Miller (Ed.), *The handbook of linguistics* (pp. 563-581). Oxford: Blackwell.

Finegan. E., & Biber, D. (2001). Register variation and social dialect variation: The register axiom. In P. Eckert, & J. R. Rickford (Ed.), *Style and Sociolinguistic Variation* (pp. 235-67). Cambridge: Cambridge University Press.

Fontanella de Weinberg, M. B. (1992). *El español de América*. Madrid: Mapfre.

Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. London, New York: Continuum Press.

Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hawkins, J. (1999). Processing complexity and filler-gap dependencies across grammars. *Language*, 75, 224-285.

Hawkins, J. (2000). The relative order of prepositional phrases in English: going beyond manner-place-time". *Language Variation and Change*, 11, 231-266.

Hawkins, J. (2001). Why are categories adjacent?. *Journal of linguistics*, 37, 1-34.

Hawkins, J. (2003). Efficiency and complexity in grammars: Three general principles. In J. Moore, & M. Polinsky (Ed.), *The nature of explanation in linguistic theory* (pp. 121-152). Stanford: CSLI Publications.

Kovacci, O. (1999). El adverbio. In I. Bosque, & V. Demonte (Ed.), *Gramática descriptiva de la lengua española* (Vol. I) (pp. 705-786). Madrid: Espasa.

Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Lavandera, B. (1975). *Linguistic structure and sociolinguistic conditioning in the use of verbal endings in 'SI' clauses*. Unpublished doctoral dissertation, University of Pennsylvania.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126-1177

Lipski, J.M. (1994). *Latin American Spanish*. London: Longman.

Lohse, B., Hawkins J., & Wasow, T. (2004). Processing domains in English verb-particle constructions. *Language*, 80, 238-261.

Marcos Marín, F. (1995). Las industrias del idioma. Posibilidades de la tecnología lingüística. *El horizonte de la sociedad de la información*, *suplemento TELOS*, *Cuadernos de Comunicación, Tecnología y Sociedad*, 41, 38-47.

Mayoral Hernández, R. (2004). Importance of weight and argumenthood on the ordering of adverbial expressions. In V. Chand, A. Kelleher, A. J. Rodríguez, and B. Schmeiser (Ed.), *Proceedings of the 23rd West Coast Conference on Formal Linguistics* (pp. 569-582). Somerville: Cascadilla Press.

Mayoral Hernández, R. (2008). A typological approach to the ordering of adverbials: Weight, argumenthood and EPP. *International Journal of Basque Linguistics*, 39, 141-159

Mayoral Hernández, R., & Alcázar, A. (2007). Diachronic changes in the position of frequency adverbials in Modern Spanish. Poster presented at *New Ways of Analyzing Variation 36*, University of Pennsylvania, Philadelphia.

Mayoral Hernández, R., & Alcázar, A. (2008). A diachronic analysis of frequency adverbials: Variation in Peninsular and Latin American Spanish. *Selected proceedings of the 4th Workshop on Spanish Sociolinguistics* (pp. 81-90). Somerville, MA: Cascadilla Proceedings Project.

Morales, A. (1986). *Gramáticas en Contacto: Análisis Sintácticos Sobre el Español de Puerto Rico*. Madrid: Playor

Moreno Fernández, F. (1993). *División Dialectal del Español en América*. Alcalá de Henares: Universidad de Alcalá.

Ranson, D. L. (1991). Person Marking in the Wake of /s/ Deletion in Andalusian Spanish. *Language Variation and Change*, 3, 133-152.

Real Academia Española: Banco de datos (CREA) [online]. *Corpus de referencia del español actual.* http://www.rae.es

Real Academia Española: Banco de datos (CORDE) [online]. *Corpus diacrónico del español*. http://www.rae.es

Silva-Corvalán, C. (1989). *Sociolingüística: teoría y análisis*. Madrid: Alhambra.

Silva-Corvalán, C. (1997). Variación sintáctica en el discurso oral: Problemas metodológicos. In F. Moreno Fernández (Ed.), *Trabajos de Sociolinguística Hispánica* (pp. 115-135). Alcalá de Henares: Universidad de Alcalá, Servicio de Publicaciones.

Silva-Corvalán, C. (2001). *Sociolingüística y pragmática del español.* Washington DC: Georgetown University Press.

Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Wasow, T., & Arnold, J. (2003). Post-verbal constituent ordering in English. In G. Rohdenburg, & B. Mondorf (Ed.), *Determinants of grammatical variation in English* (pp.119-154). Berlin: Mouton de Gruyter.

Selected Proceedings of the 11th Hispanic Linguistics Symposium

edited by Joseph Collentine, Maryellen García, Barbara Lafford, and Francisco Marcos Marín

Cascadilla Proceedings Project Somerville, MA 2009

Copyright information

ISBN 978-1-57473-432-4 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

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Mayoral Hernández, Roberto and Asier Alcázar. 2009. Technological Applications to Linguistic Research: A Corpus Analysis of Frequency Adverbials. In *Selected Proceedings of the 11th Hispanic Linguistics Symposium*, ed. Joseph Collentine et al., 242-253. Somerville, MA: Cascadilla Proceedings Project.

or:

Mayoral Hernández, Roberto and Asier Alcázar. 2009. Technological Applications to Linguistic Research: A Corpus Analysis of Frequency Adverbials. In *Selected Proceedings of the 11th Hispanic Linguistics Symposium*, ed. Joseph Collentine et al., 242-253. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2217.