

The role of lexical frequency in syntactic variability:  
Variable subject personal pronoun expression in Spanish

Daniel Erker  
Boston University

Gregory R. Guy  
New York University

Email addresses:

[danny.erker@gmail.com](mailto:danny.erker@gmail.com)

[gregory.guy@nyu.edu](mailto:gregory.guy@nyu.edu)

Mailing address for Daniel Erker:  
Boston University  
Department of Romance Studies  
718 Commonwealth Avenue  
Boston, MA 02215

Mailing address for Gregory Guy:  
New York University  
Department of Linguistics  
10 Washington Place  
New York, NY 10003

The role of lexical frequency in syntactic variability:  
Variable subject personal pronoun expression in Spanish

## Abstract.

Much recent work argues that lexical frequency plays a central explanatory role in linguistic theory, but the status, predicted effects, and methodological treatment of frequency are controversial, especially so in the less-investigated area of syntactic variation. This paper addresses these issues in a case study of lexical frequency effects on variable subject personal pronoun (SPP) expression in Spanish. Prior studies of Spanish SPP use revealed significant constraints including formal and semantic properties of the verb, and discourse factors such as a switch-reference. These appear to be confirmed in our analysis of 4,916 verbs from a spoken corpus of Spanish, along with a powerful role for lexical frequency. But the frequency effect – best configured as a discrete rather than continuous variable - is complex; statistically, it has no independent direct effect, but operates entirely through interaction with other constraints. All other constraints on SPP use are amplified in high frequency forms, and some disappear at low frequencies. Frequency thus acts as a ‘gatekeeper’ and potentiator: above some frequency threshold, significant linguistic constraints on SPP use emerge; below the threshold they do not. We propose that this reflects experience and acquisition: speakers cannot formulate hypotheses about individual lexical items until they have sufficient evidence; the threshold is the level at which speakers have enough experience with a form to do so.

These results have important theoretical and methodological implications. They require rich lexical representations incorporating frequency and collocational information. Methodologically they indicate the need for careful quantitative explorations of frequency, because its role as an enabler of other constraints produces unstable statistical results.

Keywords: Lexical Frequency, Spanish, Subject Pronouns

\*Acknowledgement footnote: The authors wish to thank Ricardo Otheguy and Ana Celia Zentella for generously sharing both their data and insights. We also wish to thank Robert Bayley, Richard Cameron, Kimberly Geeslin, Catherine Travis, and Rena Torres Cacoullos for thoughtful critiques and suggestions for the current study. We also wish to thank the referees, the editor, and the associate editor of *Language* for their insightful reviews.

Daniel Erker and Gregory R. Guy  
*Boston University & New York University*

*Information is pretty thin stuff unless mixed with experience.*

*-Clarence Day*

**1. INTRODUCTION.** In the continuing search for explanatory principles, contemporary linguistic inquiry has seen a range of work that attributes central importance to the frequency of occurrence of linguistic items in regular usage. The basic concept of this work is that words or linguistic structures that speakers use often will have distinctive mental representations, and will be treated differently in language processing and linguistic variation and change compared to forms that are relatively rare. Items that are highly practiced and very familiar will be recognized quicker, articulated more easily, changed more or less readily, perceived as more grammatical, and accorded distinctive mental status; in effect, practice makes perfect, or at least, practice makes different. One of the strongest exponents of this position is Bybee, who suggests:

A new theory of language is emerging as a convergence from many different research paradigms. Evidence from phonology rounds out the theory, demonstrating that every aspect of language can profitably be re-examined in light of the important frequency effects. (2002: 220)

The different research paradigms that Bybee cites include: language change, where frequent words are found to lead certain sound changes (cf. for example, Phillips 1984, 1999; Bybee 2001, 2002); psycholinguistics, where lexical frequency affects reaction times in word recognition (Vitevitch 1998 and 2002); laboratory phonology (Pierrehumbert 2001, Beckman and Pierrehumbert 2003b, Guy, Hay and Walker 2008); language acquisition (Ellis 2002a, 2002b); and linguistic variation (Bybee 2002, Myers and Guy 1997, Erker 2011, File-Muriel 2011).

Despite this widespread interest in frequency, however, there is no clear consensus on the nature and direction of the effects of frequency on linguistic processes. Indeed, there are claims that frequency has different effects on different kinds of processes; for example, higher frequency is said to accelerate phonological reduction (so that frequent forms are most affected), but low frequency is claimed to favor the expansion of general patterns, as in cases of regularization (thus infrequent forms are most affected). The mechanisms expounded for frequency effects on phonology, such as articulatory practice of oft-repeated gestures, are not necessarily homologous to those that operate in syntax, like the greater entrenchment of frequent elements. And it is not always clear how frequency is best defined – locally or globally, continuously or discretely, by lemma, form, or collocation, at what level of granularity, et cetera.

Given these unresolved issues, there remains a need for quantitative examination of the significance of lexical frequency, especially in the domain of morphosyntax. This paper seeks to illuminate some of these issues with an in-depth analysis of a case of linguistic variation. We focus on a well-understood syntactic variable in a well-known language, namely, the variable occurrence of personal pronouns as the overt subjects of tensed verbs in Spanish. The considerable body of research devoted to this topic has shown that the occurrence of overt subject personal pronouns (SPPs) is regularly and significantly conditioned by several syntactic and semantic properties of the tensed verb, such as tense/mood and person/number, but it has yet to address the question of any effect of lexical frequency. As such it provides a firm grounding for the testing of new models, and a good candidate for “reexamination in light of important frequency effects”. As we will show, in a quantitative analysis of pronoun occurrence in a corpus of Latin American Spanish speakers in New York City, frequency indeed turns out to play a significant role in the occurrence or non-occurrence of SPPs in Spanish, but in an unexpected way. Higher frequency does not by itself favor or disfavor pronoun use; rather it has significant interactions with all other constraints affecting SPP occurrence. High frequency potentiates or amplifies the effects of other constraints. Thus in a favorable context for SPP use, such as with a past imperfect verb form, higher lexical frequency of the verb is associated with higher rates of SPP use, but in a disfavoring

context, such as with a preterite verb form, higher lexical frequency is associated with lower rates of pronoun occurrence.

**1.1. THE place of frequency in linguistic theory.** Traditional theoretical models in linguistics postulate a unique abstract representation of each lexical item which has little or no place for information about lexical frequency. This tradition, dating at least from the Neogrammarians and exemplified in contemporary work by the generative tradition, treats each lexical item as a string of phonological units associated with syntactic features and a semantic representation. These abstract formal units of representation are expected to capture all relevant linguistic information about the word. That is, a word is not expected in such a model to have any existence, or unique linguistic behavior, that is not captured by the formal representation. Each word does NOT have its own history, nor syntactic idiosyncrasies, in such models. The formal components of the representation are discrete entities (segments, phonological features, phi-features, etc.), not gradient or scalar properties like lexical frequency.

In opposition to this view there have emerged in recent decades several perspectives that emphasize continuous and gradient properties in the linguistic system. Prominent among these are ‘usage-based’ and ‘exemplar-theoretic’ models (Bybee 2001; Pierrehumbert 2001, Bybee and Torres Cacoullos 2008), which postulate that speakers store extensive, detailed memories of the utterances and words they encounter in language use, and utilize these memory sets to derive generalizations and to generate production targets when they are speaking. In such models, speakers implicitly retain massive information about lexical frequency and contexts of use.

Usage based models further postulate that collocational information is retained in and deducible from memories of linguistic experience. The principle device available in a traditional generative model for treating frequently occurring associations between pairs or groups of words is to represent them as if they are enlarged lexical items; thus idiomatic expressions such as *lend a hand* and *kick the bucket* are not parsed as phrases, but as a type of compound lexical entry with a single meaning (respectively ‘help’ and ‘die’.) Remembered exemplar clouds, however, may in principle allow speakers to deduce what words are likely to co-occur, with what probability, formulating

probabilistic generalizations well-beyond the level of the fixed phrase or idiom. This permits a scalar view of collocations, with lexicalized idioms merely being one extreme case of highly frequent co-occurring items. This is the kind of information that is relevant to the present study; if specific verb forms show distinctive associations with rates of occurrence of overt subject pronouns, whether directly, or indirectly through their scalar level of lexical frequency, this would imply that speakers' knowledge of language incorporates collocational information.

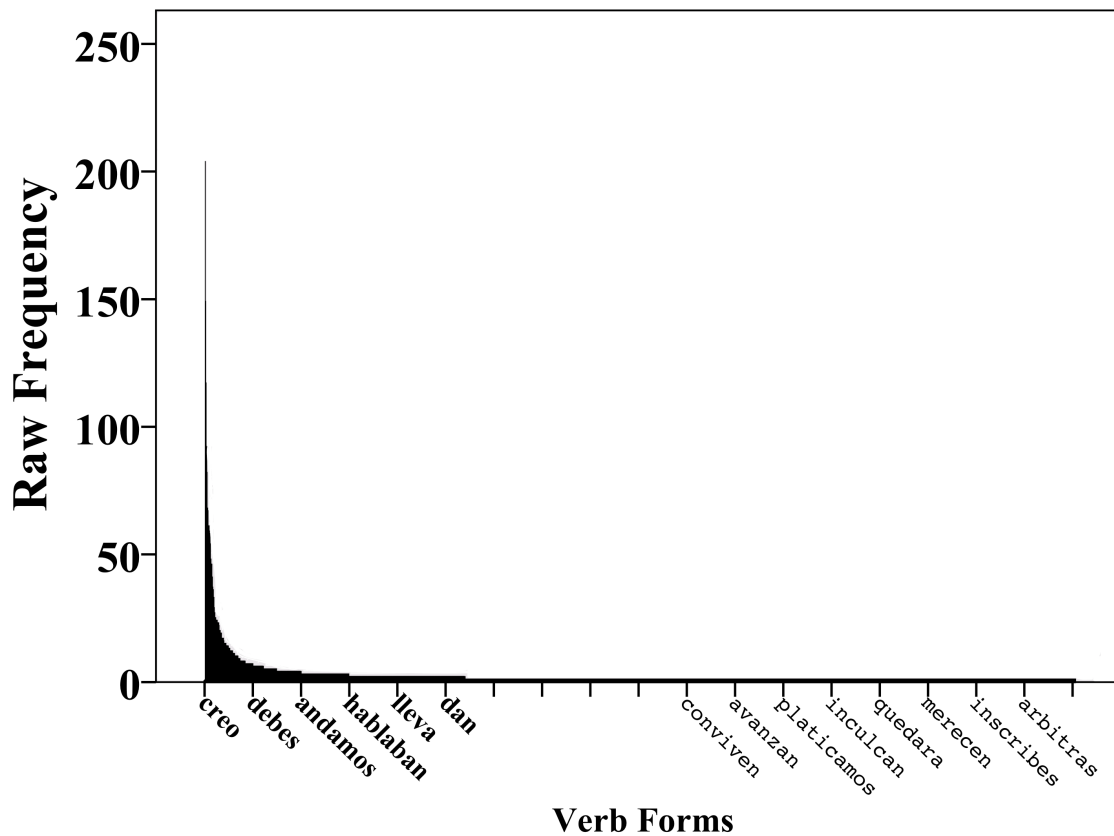
The predictions that usage-based models draw from this postulated general availability to speakers of detailed frequency information about lexical items and collocations point in several different directions. Much of the work in these approaches has focused on phonological implications, where frequent repetition of forms is predicted to favor phonetic REDUCTION or LENITION (cf. Bybee 2010); thus the extremely frequent English word *and* has an exceptionally high rate of deletion of the final /d/, compared to less frequently occurring words like *sand*, *band* (cf. Guy, Hay and Walker 2008). Since final /t,d/ deletion appears to be a general lenition process of English, this suggests overapplication of the process in high frequency forms. However, in morphological processes, especially diachronic ones, high lexical frequency is often associated with underapplication of general tendencies. Thus the long-term drift in English verbs from strong to weak preterites (replacing *clomb* with *climbed*, *holp* with *helped*, etc.) has been resisted by many of the highest frequency strong verbs, such as *be*, *eat*, *see*. In this case, high frequency is associated with a CONSERVING or ENTRENCHMENT effect (Bybee and Thompson 1997, Bybee 2010). This effect has been characterized in terms of high token frequency combined with low type frequency (cf. Travis & Silveira 2009, Poplack 2001): the surviving strong verbs are few in number (few types) but have many tokens, whereas the spreading weak verb morphology includes many more types, of which many have low token frequencies. In more purely syntactic cases, the Conserving effect is also claimed for lexically-specific exceptions to syntactic changes. Thus Bybee and Thompson (1997) suggest that the high frequency of English modal verbs is the reason why they retain some Middle English syntactic possibilities that were formerly available to all verbs, but have been lost in ordinary main verbs, such as inversion with the subject (*Can you see it?* vs. *\*See you it?*) and post-posed negatives (*They might not go* vs. *\*They go not*). Finally,

the postulated retention of collocational information leads Bybee to argue for a third effect of high frequency co-occurrence: the AUTONOMY effect, which refers to the emergence of ‘chunking’ of word sequences or phrases that begin to be accessed on some level as (relatively autonomous) units (cf. Bybee 2010). In the phonetic domain, such chunking is argued by Bybee to further favor lenition, especially in cases of grammaticalization (e.g., *going to* > *gonna*). An additional effect of collocation is the apparent generalization of the effects of frequently occurring contexts to the lexical items that occur in them (cf. Bybee 2003). Thus Guy et al. 2008 find that words that most commonly occur in favorable contexts for final /t,d/ deletion have higher deletion rates even when not occurring in those context (see similar findings in Bybee 2002 and Brown 2004).

To summarize, the generative tradition and its predecessors postulate no lexical representation of frequency, while the usage-based, exemplarist approaches postulate exhaustive representations of all encountered exemplars, and hence unlimited information about lexical frequency and contexts of use. It is of course possible to entertain models that are intermediate between these two. Such a hybrid model might, for example, incorporate abstract representations that are enriched with parameters for frequency, collocational information, activation level, or other lexically specific traits, which are updated with experience, without entertaining comprehensive retention of prior linguistic experiences. An informed choice among these possible theoretical positions depends, of course, on many considerations beyond the scope of this paper. Our approach will be empirical: we investigate whether and how frequency affects Spanish pronoun use, and consider the implications of our results for the various effects of lexical frequency.

**1.2. METHODOLOGICAL challenges in the study of frequency.** The study of lexical frequency raises a number of methodological and analytical issues. First, when it comes to collecting data, frequency studies are constrained by the fact that the frequency distribution of words follows Zipf’s Law, a power law distribution: there are very few words with very high frequencies and many words with very low frequencies. Figure 1 illustrates this distribution in the data for the current study.

Figure 1. Zipf's Law: Raw frequency distribution of Spanish verb forms



This distribution imposes great practical complications for data collection. In any corpus of natural speech, each low frequency lexical item is represented by very few tokens; since the reliability of quantitative analysis is sensitive to sample size, very large amounts of data have to be processed in order to get enough data for reliable statistical analysis of individual infrequent words.

A second methodological challenge is deciding what and how to count. There are a variety of options in the quantification of frequency. For example, should frequency be measured in terms of lemmas or in terms of surface word-forms – for instance, does one count *think*, *thinks*, and *thought* as three different lexical items or take them all together as instances of the verb ‘to think’? Given the primacy accorded to surface forms in usage-based models, we focus on surface forms rather than lemmas or paradigms. Next, should frequency be measured locally or globally? Frequency studies often use counts from large corpora made up of heterogeneous texts, such as the Brown Corpus or Celex,

but how relevant are these to the usage of a specific speech community or individual? An alternative is to use local frequency counts generated on the basis of speech taken from only the community or individual, but these raise sampling issues: a few hours of recordings do not give a fine-grained picture of an individual's lexical usage. Good arguments can be made for several such approaches, but with different analytical consequences. Since the spirit of usage-based models is to attribute 'grammatical' patterns to speakers' experiences, we favored a measure that approximates the usage prevailing in the local speech community, namely frequency counts in our own corpus.

A third methodological issue in the study of frequency relates to how it is operationalized in quantitative analyses. Specifically, should frequency be characterized continuously in terms of counts, ranks, or occurrences per volume of words? Or, alternatively, should it be considered discretely, opposing frequent and infrequent forms? If the latter, what are the best criteria for segmenting a set of observations into frequency bins? Continuous treatments ordinarily suppose a monotonic, and typically linear effect: each incremental change in frequency predicts some corresponding increment or decrement in the dependent variable. But many other relationships are possible, such as step or threshold functions, which are better analysed by binning, and curvilinear functions (such as exponential and logistic). Given the present limited understanding of frequency effects in linguistics, it is not always clear which of these analytical options should be adopted. In the present study we explore several alternatives.

Finally, there is the question of how lexical frequency relates to other linguistic or social factors that may condition variable phenomena. In variationist sociolinguistics, the analysis of a variable linguistic feature proceeds by establishing the phonological, morphological, syntactic, semantic, and discursive factors that significantly affect the choice of variants. The social distribution of forms is investigated by examining the social and personal characteristics of speakers and their preferences for particular variants. These linguistic and socio-personal predictors are seen as exerting differing degrees of influence on, or differing strengths of association with the dependent variable (Bayley 2002a). The ranking, weight, and significance of predictor variables is regarded by many scholars as constituting the variable component of the grammar of the language (Cedergren & Sankoff 1974, Poplack 2000, Guy 2005, 2007).

In such a framework, the status of lexical frequency is not obvious. The simplest approach is to treat lexical frequency as another linguistic predictor (or ‘factor group’) in the conventional sense. That is, when frequency is investigated as a constraint on variation, it is often assumed to have similar status to other linguistic predictors. Thus:

This article only addresses word frequency as a conditioning factor in phonological variation and assumes that the variation remaining after frequency is considered is due to other well-established social and linguistic factors. My goal is to demonstrate that word frequency and context frequency are factors that can affect variation and should be taken into account in the future studies of phonological variation and change.  
(Bybee 2002: 262:)

The soundness of this perspective – that frequency is a conditioning factor on par with the other kinds of factors routinely studied in work on linguistic variation – is relatively untested. Given claims made about frequency in the frameworks that assert its centrality, there is reason to doubt that it operates as just another constraint orthogonal to all the others. Consider for example ‘other well-established linguistic factors’ that reflect abstract representations and rules; since usage-based models, including Bybee’s, often deny any primary role or existence to such abstractions, treating them instead as analogic generalizations derived from a remembered corpus of lexical exemplars, one might reasonably expect them to disappear once the frequency and identity of individual lexical items are accounted for. Indeed, any model that asserts the primacy of words over structural accounts (exemplar theory, lexical diffusion, etc.) effectively claims that words cannot be treated as assemblages of phonemes, morphosyntactic categories, semantic properties, et cetera.; this implies that the relationship between a word-based and a structurally-based account is non-linear, and potentially chaotic. Methodologically, this would imply that incorporating frequency into a quantitative study might well produce interactions, unstable results, and other statistical disruptions. Furthermore, frequency will necessarily have strong colinearities with many potential linguistic factors: function

words have extremely high frequencies compared to content words; morphologically irregular forms are notorious for being high in frequency; unmarked structures will have higher frequencies than marked forms. Such relationships complicate quantitative analysis. Therefore, it is unwise to simply add lexical frequency into an established set of linguistic predictors for a given variable phenomenon without examining the implications. It may be the case that frequency has a different relationship with both variable phenomena and the kinds of factors typically understood as conditioning them.

For the syntactic variable studied here – the variable occurrence of subject pronouns with verbs of differing lexical frequencies – usage-based models suggest several possible, sometimes contradictory, hypotheses about frequency effects. The Reduction effect of frequency, although elaborated for phonetic/phonological reduction, has some parallels in syntactic variation: Jaeger (2006) finds that in English verb + complement clause constructions, higher frequency verbs have more complementizer omission, and in relative clauses, higher frequency antecedent NPs have more relative pronoun omission. In our case this would imply fewer subject pronouns with high frequency verb forms.

The Conserving effect of frequency, however, has less clear implications for the SPP case. It has been formulated principally in terms of apparent resistance to changes involving generalization. We have no evidence of ongoing change in Spanish SPP use. Nevertheless, given the comparative evidence of change in related Romance languages towards SPP occurrence rates that are elevated (e.g., Brazilian Portuguese) or categorical (e.g., French), and the higher rates of occurrence in Latin American Spanish than in Spain, it is likely that the historical direction of change has been from lower to higher rates of overt SPP occurrence. From this perspective lower SPP rates are archaic, so high frequency verbs should Conserve a lower incidence of SPPs. This would be analogous to Poplack's finding (1992, 1995) that a few high-frequency matrix verbs in French (especially *falloir*) continue to trigger high rates of usage of subjunctive verb forms in embedded clauses, despite a general historical trend towards loss of the subjunctive – in other words, high lexical frequency in a triggering verb is associated with an archaic rate of usage of a dependent form, defying a change process.

Finally, the usage-based prediction that may appear most relevant to our case is the Autonomy effect: frequent collocations begin to be accessed or parsed as units. This would suggest that particular high frequency pairings of pronoun with verb are treated as ‘chunks’ and conserved in that relationship. But as it has been formulated, Autonomy makes equally valid predictions for both the absence and the presence of an overt pronoun: a frequent verb form that commonly occurs with a zero pronoun would become entrenched with the zero collocation, while a form that commonly occurs with an overt SPP would become entrenched in that association. This does not help us predict which frequent verb forms should have the one or the other, although it could suggest a prediction that discourse-level effects, such as inter-clausal continuity of reference, might be weaker in high frequency forms if the latter are treated as chunks. In sum, therefore, the general tendencies identified by usage-based models yield two predictions that high lexical frequency should be associated with low SPP rates (Reduction and Conservation of older patterns), and one tautological prediction, that high frequency items should retain whatever tendencies to co-occur with SPPs that they happen to exhibit, whether high or low. None of these general tendencies implies a systematically high rate of overt SPP occurrence for high frequency verbs.

**1.3. CURRENT study of Spanish Subject Pronouns.** The range and productivity of prior research on Spanish subject pronoun use make this variable an excellent site for exploring questions of lexical frequency. Indeed, few variable linguistic phenomena have been so widely investigated as Spanish SPPs, which have been examined in speech communities throughout Latin America (Avila-Jiménez 1995, Barrenechea & Alonso 1977, Bentivoglio 1987, Cameron 1993, 1994, 1995, 1996, Lipski 1994, 1996, Morales 1997, Miyajima 2000, Travis 2005), the Iberian Peninsula (Cameron 1993, Davidson 1996, Enríquez 1984), and the United States (Bayley & Pease-Alvarez 1996 and 1997, Bayley 2002b, Cameron and Flores-Ferrán 2004, Flores-Ferrán 2004, Lapidus and Otheguy 2005a, 2005b, Otheguy, Zentella, and Livert 2007, Silva-Corvalán 1982, 1994, 1997a, 1997b). SPP expression is in several respects a classic linguistic variable. First, it is a discrete, binary phenomenon that has little bearing on the truth conditions of utterances. Thus the utterances in 1 below are equivalent in the sense that they have

identical logical form and truth conditions.

- (1) a. *Yo hablo.*  
       'I speak'  
       b. *Hablo.*  
       'I speak'

Secondly, the presence or absence of pronouns such as *yo* is clearly conditioned by both linguistic as well as extralinguistic factors (Bayley & Pease-Alvarez 1996, Cameron 1993, Guitart 1982, Otheguy, Zentella, Livert 2007, Silva-Corvalán 1982). Thirdly, while some of the conditioning factors are broadly stable across all Spanish dialects, the relative strengths of others have been shown to vary intriguingly across and within communities (Barrenechea & Alonso 1977, Bentivoglio 1987, Enríquez 1984, Lipski 1994). Hence SPPs have provided a valuable site for conducting research on a range of topics, including dialectology, sociolinguistic variation, and language contact and change. They have also been studied extensively in the context of Spanish second-language acquisition (Liceras 1989, Lafond, Hayes and Bhat 2000, Geeslin and Guijarro-Fuentes 2006, Montrul and Rodriguez Louro 2006, Geeslin and Gudmestad 2011).

Recently in *Language*, Otheguy, Zentella, and Livert (2007) examined subject pronoun use in Spanish spoken in New York City. The current study draws its data from the same source, *The Otheguy-Zentella corpus of Spanish in New York City* (described in section 2 below). Among their valuable contributions Otheguy et al. tested the effects of a large set of linguistic factors that have been hypothesized to condition pronoun use<sup>1</sup>, finding significant effects for a number of features of the tensed verb: person/number, tense/mood/aspect (TMA) and semantic content. In addition, they found a significant discourse-level effect of continuity of reference across clauses, consistent with previous work on switch-reference. This study also investigated social variation in SPP use, finding differences in the effects of these constraint between various nationality groups of Spanish speakers; they further showed that these differences diminish among Spanish speakers born and raised in NYC compared to those recently arrived from Latin America. They interpret this to be a consequence of language and dialect contact.

Given this depth of prior research, we are well positioned to ask whether lexical frequency plays a role in pronoun use. A well-established set of linguistic conditioning factors accounts for much of the variability in Spanish SPP use; we now examine what additional contribution is made by lexical frequency. This study also serves to address the relative dearth in the frequency literature of studies of variable syntactic phenomena. Thus SPP use provides a promising site for responding to Bybee's call to 'reexamine' additional aspects of language for frequency effects.

**2. DATA and method.** The data for this study consist of 4,916 Spanish verb forms taken from the *Otheguy-Zentella corpus of New York City Spanish*. The corpus is a collection of sociolinguistic interviews with 141 Spanish speakers living in New York City, drawn from the six largest national populations in the city: Colombians, Cubans, Dominicans, Ecuadorians, Mexicans, and Puerto Ricans. From this corpus, 12 were chosen for the current study, six of Mexican and six of Dominican origin. In each national group, half of the speakers selected were recent arrivals to NYC and half were lifelong New Yorkers. We chose these speakers because Mexican and Dominican varieties of Spanish are typically very different with respect to overall rates of pronoun use. Spanish in the Dominican Republic is characterized by relatively high rates of overt pronouns while the Spanish spoken in most of Mexico typically shows low rates of pronoun use (Lipski 1994). Given that these two varieties occupy the poles of pronominal behavior, they are good candidates for asking whether the role of lexical frequency may vary dialectally.

Each of the 4,916 verbs either occurred with a subject pronoun (*tú sabes* 'you know') or could have but did not occur with a pronoun (*sabes*). Of the 4,916 verbs included in the study, 1,709 (34.8%) occurred with overt pronouns while 3,207 (65.2%) were null pronominally. This is consistent with the results of Otheguy et al., who reported 33.4% overt pronoun occurrence.

Two points are important to note about the relationship between the present data set and that analyzed in Otheguy *et al.* 2007. That study did not address lexical frequency, nor examine individual verb forms. Consequently the present study required a new pass through the corpus to identify specific lexical items and analyze their

frequency. Second, the substantially larger data quantity of Otheguy *et al.* was motivated by their objective of characterizing an entire speech community encompassing a broad array of social variables (six national dialects, speakers stratified by generation of migration, age of arrival, duration of stay, language competence in English, etc.), which made it necessary to look at many speakers. Given the focus in the present paper on lexical frequency and its interaction with linguistic constraints, the N of almost 5,000 clauses drawn from 12 speakers provides a statistically reliable sample for testing the predictions and effects we address. As we will show, the main result reported here – the amplification of linguistic constraint effects in high frequency forms – is systematically replicated in various subsets of the data, which confirms both the finding and the adequacy of the sample.

**2.1. LINGUISTIC constraints.** As noted above, previous work on variable SPP use in Spanish has consistently shown it to be constrained by several linguistic factors. In this section we outline the variables that we investigated. For purposes of comparability, we adopt the same constraint definitions as Otheguy *et al.* (2007) wherever possible. From their constraint set we examined four that consistently had significant effects: person/number, tense/mood/aspect, and semantic content of the tensed verb, and continuity of reference between clauses. To their set of constraints we have added two that were not addressed in their study: *morphological regularity*<sup>2</sup> and *lexical frequency*. In what follows we describe the values of each constraint investigated.

*Morphological regularity* – Paradigmatic regularity of verb forms was investigated for this study because of two typical associations of irregular forms. First, they are notoriously concentrated among the more frequent verbs, and hence must be controlled for to avoid a potential confound – that is, to test whether any frequency effect of verb form on SPP use is independent of the regular-irregular opposition. Second, some effects on SPP use have been claimed to be functionally motivated: it is argued that the higher saliency of the first plural inflections (with the distinctive *-mos* suffix) and the preterite forms (where all six person/number categories have distinct inflections) make overt SPPs communicatively redundant, motivating lower SPP frequency in these

contexts. Since irregular forms are highly salient, they offer an additional perspective on this issue.

Verb forms were classified for this variable based on the relationship between surface morphology and the associated root. For example, *tengo* and *tenía* were coded on the basis of their relationship to *tener*. Verb forms were considered irregular if they met any of three criteria; (1) their derivation involved velar insertion, *tener* → *tengo* ‘I have’, *hacer* → *hagas* ‘you make’; (2) their derivation involved velar insertion plus a change in vowel quality, *decir* → *digo* ‘I say’, *traer* → *traigo* ‘I bring’; or (3) the form was unique, that is, not derivable from a root, for example, *eres* ‘you are’, *fue* ‘she went’, *sepan* ‘they know’ (from *ser*, *ir*, *saber* respectively). In total, 3,170 forms were coded as regular and 1,682 were coded as irregular<sup>3</sup>.

*Person and Number* – Verb forms were coded for one of six person-number values: *first-*, *second-*, and *third-singular* (*hablo*, *hablas*, *habla* ‘I, you, she speak(s)’), and *first-*, *second-*, and *third-plural* (*hablamos*, *hablais*, *hablan* ‘We, you, they speak’). Only a tiny fraction of the forms in the study, 8 in total, were *second-plural*; these are not considered in the quantitative analysis below. By contrast, nearly half – 2,456 verb forms – were *first-singular*. The distribution among other forms was: 846 *second-singular*, 702 *third-singular*, 305 *first-plural*, and 605 *third-plural* forms.

*Tense-Mood-Aspect* – The study distinguished 11 different tense-mood-aspect (TMA) combinations: *indicative present*, *preterite*, *imperfect*, *perfect*, and *future* (e.g., *habla*, *habló*, *hablaba*, *ha hablado*, *hablará*, respectively): *subjunctive present*, *past*, and *perfect*; (*hable*, *hablara*, *hubiera hablado*) and also the *periphrastic future*, *imperative*, and *conditional* (*va a hablar*, *hable*, *hablaría*). The majority of the study’s verb forms were the *indicative present* (2,695), *preterite* (877), or *imperfect* (708). The next largest groups were the *perfect-indicative* and the *periphrastic future*, with 176, and 140 forms, respectively. The various *subjunctive* forms occurred infrequently, and were grouped together with 164 tokens. The *imperative*, with 110 forms, the *conditional*, with 29, and the *future-indicative*, with 15, were the least frequent TMA combinations in the data.

*Semantic Content* – Verb forms were assigned to one of three semantic classes: *mental activity*, *stative*, or *external activity*. This represents a minor departure from the coding of Otheguy et al., which was largely based on that of Enriquez (1984). Those

studies distinguished from the general *mental activity* class (verbs such as *pensar* ‘to think’, *comprender* ‘to understand’, *imaginar* ‘to imagine’) a separate *estimative* class of verbs that offer opinions or express judgements (e.g., *admirar* ‘to admire’, *creer* ‘to believe’, *desear* ‘to wish’). Both of these classes were found in those studies to be associated with high pronoun rates. Enriquez argues that the presence of a pronoun tends to individualize the subject, and/or focus that subject as opposed to other possible subjects. This individuation is more likely to occur with mental activity and estimative verbs, which individualize the opinion or sentiment of the subject, or imply greater subjectivity. These two classes combined in our data accounted for 840 forms. The stative class, comprised of verbs that reflect non-dynamic processes, such as *estar* ‘to be’, *tener* ‘to have’, and *vivir* ‘to live’, represented 1,438 forms. Lastly, 2,601 forms were coded as *external activity verbs*. This broad category includes forms that refer to a wide range of external actions, such as *comprar* ‘to buy’, *escribir* ‘to write’, and *enseñar* ‘to teach.’

*Switch Reference* – This study made a binary distinction between *switch in referent* and *not a switch in referent*, defined in terms of the relationship between the referents of verbal subjects in adjacent clauses. Specifically, we look from the ‘target verb’ back to the ‘trigger verb’, which is the immediately preceding finite verb. If the subjects of both verbs have the same referent, then there is *not* a switch in referent, otherwise there is a switch. Of the verbs in this study, 2,653 represented a switch in referent, 2,233 did not (the remaining 30 tokens were unclear because of overlapping utterances or indistinct recording). This criterion ignored changes in TMA or speaker. For instance, continuity of reference is maintained in a sentence like *Cuando yo tenía veinte años, fui a España para estudiar* ‘When I was 20 years old I went to Spain to study’, despite differences in TMA between the target (*fui* ‘I went’ *1sg, preterite indicative*) and the trigger (*tenía* ‘I had’ – *1sg, imperfect indicative*). This is also true of the following example, in which trigger and target are spoken by different individuals. Though not in the same person (*2sg* vs. *1sg*) they are nonetheless co-referential:

Interviewer: *¿Tienes muchos amigos aquí?* ‘Do **you** have many friends here?’

Interviewee: *Sí ya tengo mas amigos de los que tenía allá.* ‘Yes, **I** have more here than I did there.’

By contrast, however, the following text shows a switch in reference between trigger and target in the last clause, signaled by an overt pronoun: *Yo limpiaba mi habitación, o sea ayudaba siempre en lo que ella necesitaba.* ‘**I** would clean my room, which is to say, **(I)** always used to help out with whatever **she** needed.’

**2.2. LEXICAL frequency.** In addition to the linguistic variables described above, all forms were also analyzed in terms of *lexical frequency*, which was measured in several ways. In order to approximate the usage of the speech community, frequency was defined locally, in terms of the frequency of occurrence of verb forms in our corpus of 4,916 verbs. Our basic frequency measure was *raw frequency*, a count of the number of times an item occurred in the corpus. For example, *creo* ‘I believe’, occurred 208 times, and *sé* ‘I know’ occurred 148 times, while *amenaza* ‘he threatens’ and *convivíamos* ‘we coexisted’ each occurred just once. Given the Zipfian distribution of frequencies, we also used a *log frequency* measure, calculated as the base-10 logarithm of the raw frequency; this useful transform improves resolution among the many low frequency items, while compressing the sparsely populated sprawl of high frequency items.

Both the *raw frequency* and *log frequency* measures are continuous in nature and hence best suited for identifying a uniform monotonic relationship between pronoun use and lexical frequency. But, as we will show, this relationship is not uniform, so neither of these measures is adequately revealing. More illuminating was a measure of *discrete frequency* distinguishing two groups: *frequent* and *infrequent* verb forms. Frequent forms were defined as those that individually constituted at least 1 percent of the corpus. Thirteen forms met this criterion. Together, they comprised 22.8 percent of the corpus, or 1,120 forms, as shown in Table 1.

Form	Raw Frequency	% Of the corpus
<i>creo</i> 'I believe'	204	4.1
<i>sé</i> 'I know'	148	3.0
<i>digo</i> 'I say'	117	2.4
<i>tengo</i> 'I have'	92	1.9
<i>sabes</i> 'you know'	82	1.7
<i>ves</i> 'you see'	68	1.4
<i>estaba</i> * <sup>4</sup> 'I/he/she/you was/were'	67	1.4
<i>estoy</i> 'I am'	61	1.2
<i>tenía</i> * 'I/he/she/you had'	61	1.2
<i>era</i> * 'I/he/she/you was/were'	59	1.2
<i>soy</i> 'I am'	58	1.2
<i>fui</i> * 'I was/I went'	54	1.1
<i>es</i> 'he/she/you are'	49	1
Totals	1,120	22.8

All other forms were considered *infrequent*. A comparison of these two groups shows that they are significantly different from each other with respect to mean raw frequency. Among frequent forms, mean raw frequency is 108.1, while for infrequent forms it is 8.1. Table 2 summarizes these data.

Discrete Frequency	Mean Raw Frequency	N Tokens
Infrequent (Verbs that are each <1% of data)	8.2	3,796
Frequent (Verbs that are each >1% of data)	108.1	1,120
t(4914) = 106 p <.001		

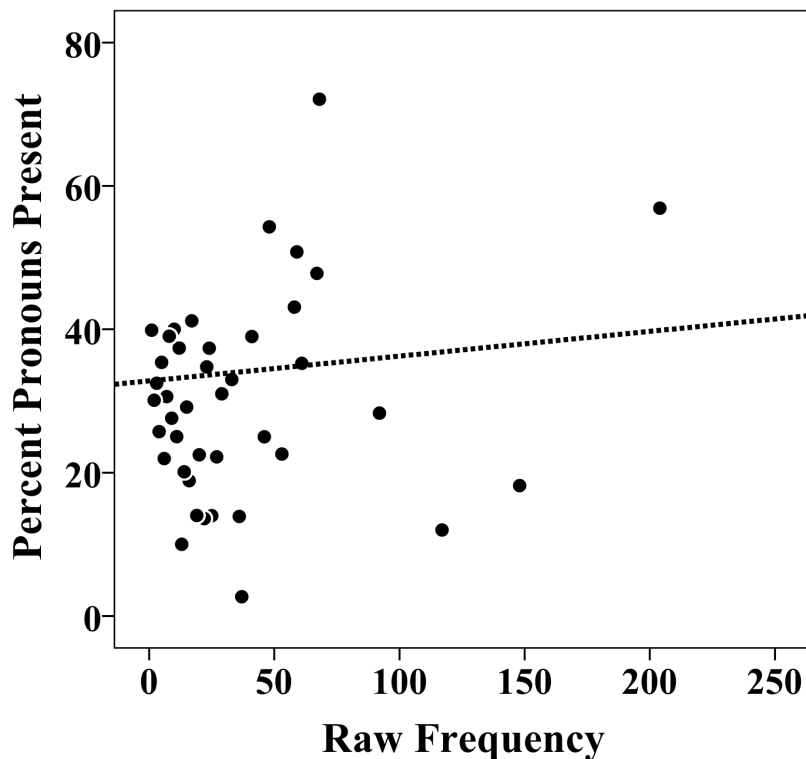
Other discrete treatments of frequency are easy to design, for example by adjusting the cut-off point, or by designating more than two frequency categories. We explored several such approaches and concluded that the one described here was the most revealing of the relationship between frequency and SPP use.

**3. RESULTS.** As indicated above, our analysis considers the linguistic constraints on SPP occurrence found by previous researchers as well as frequency effects. We begin by presenting univariate analyses for each of the constraints, confirming a main effect for

all of them, although the effects for the various frequency measures are weak and sometimes contradictory. The effect of frequency is subsequently clarified in multivariate analyses, which show consistent interaction between this variable and the other conditioning factors in the study<sup>5</sup>.

**3.1. MAIN effects of lexical frequency.** Does lexical frequency significantly affect pronoun use? The simplest approach to this question is to look for main effects of the *raw*, *log*, and *discrete frequency* measures. The first two are continuous, and allow us to investigate whether there is a linear or at least monotonic relationship between lexical frequency and pronoun use. That is, do SPP rates vary as a function of continuously defined increases or decreases in the lexical frequency of verb forms? Consider Figure 2, which plots raw frequency against percent SPPs present.

Figure 2. Raw frequency and percent SPPs present

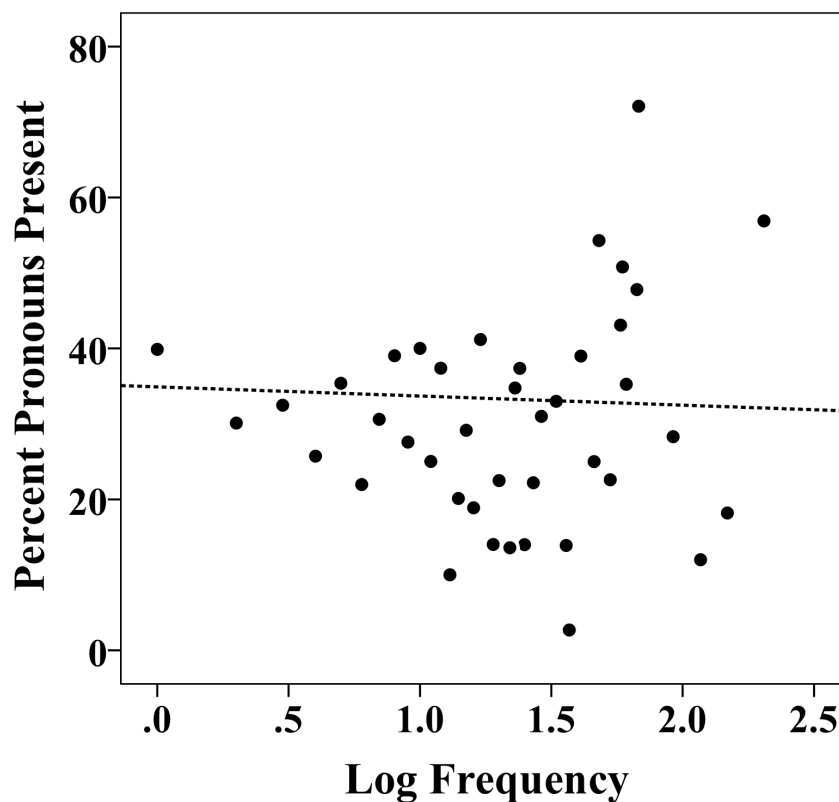


Reflecting Zipf's law (there are many rare forms but few frequent forms), most data points are concentrated in the left, low-frequency end of the figure. This makes it hard to

visually discern the presence of a linear trend, but the correlation statistic suggests a very weak but nevertheless significant association:  $r(4,916) = .057, p < .001$ . This implies that the more frequently a form occurs, the slightly more likely it is to have an overt SPP.<sup>6</sup>

At first blush, these results support the hypothesis of a relationship between pronoun use and lexical frequency. But the effect of this frequency measure is extremely weak compared to the other constraints described below. Indeed, further inspection suggests that a linear characterization of the relationship between the SPP use and lexical frequency is inadequate. Consider Figure 3 below, which uses log frequency instead of raw frequency. The logarithmic transform compensates for the Zipfian distribution and facilitates visualizing trends in the data.

Figure 3. Log frequency and percent SPPs present

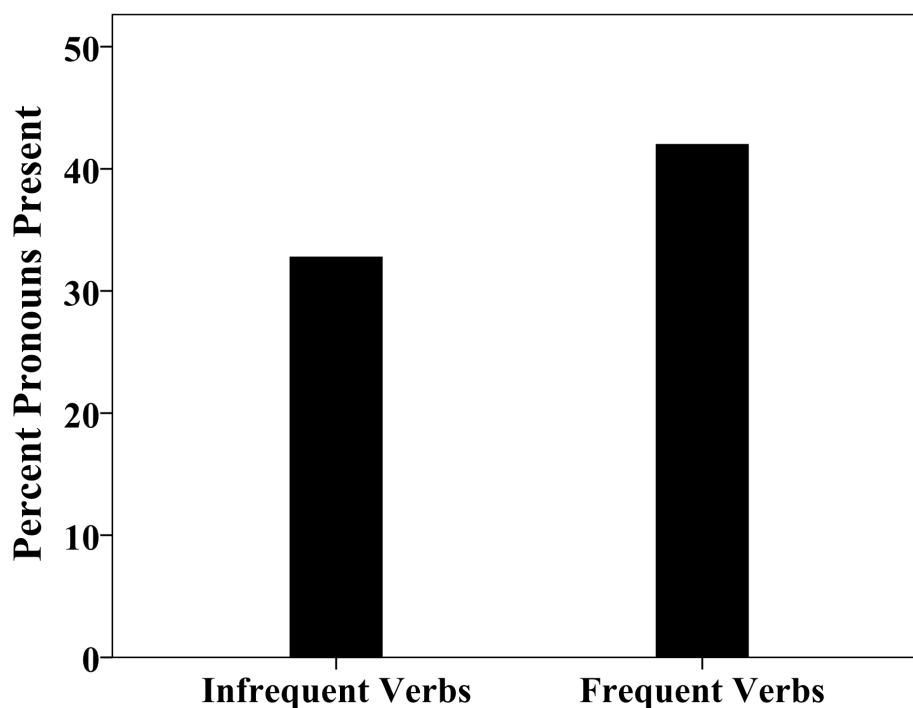


In this log-frequency plot, the correlation statistic is negative (higher frequency is associated with lower SPP use); it is also significant but very weak,  $r = -.081, p < .001$ . This is the opposite of the apparent effect of raw frequency, which tends to undermine

whatever confidence we might put in such weak correlations. What is particularly notable is that most of the data points representing higher frequency verb forms fall far from the linear trend in the data. In other words, for a large fraction of verb forms, SPP rates are poorly predicted by frequency, and the predictions do not improve with increasing frequency. This is of course contrary to what the central limit theorem predicts for errors due to random statistical fluctuation, suggesting that the differences in SPP use between the higher frequency forms are real, and that SPP rates may be diverging with increased frequency, rather than converging.

Our remaining frequency measure, *discrete frequency*, offers another perspective on the relationship between frequency and SPP use. As Figure 4 shows, discretely defined lexical frequency significantly affects SPP use: frequent forms occur with SPPs at a significantly higher rate, 42%, than do infrequent forms, 33% ( $t = 5.6$ ,  $p < .001$ ).

Figure 4. Infrequent and Frequent forms by Percent SPPs Present.

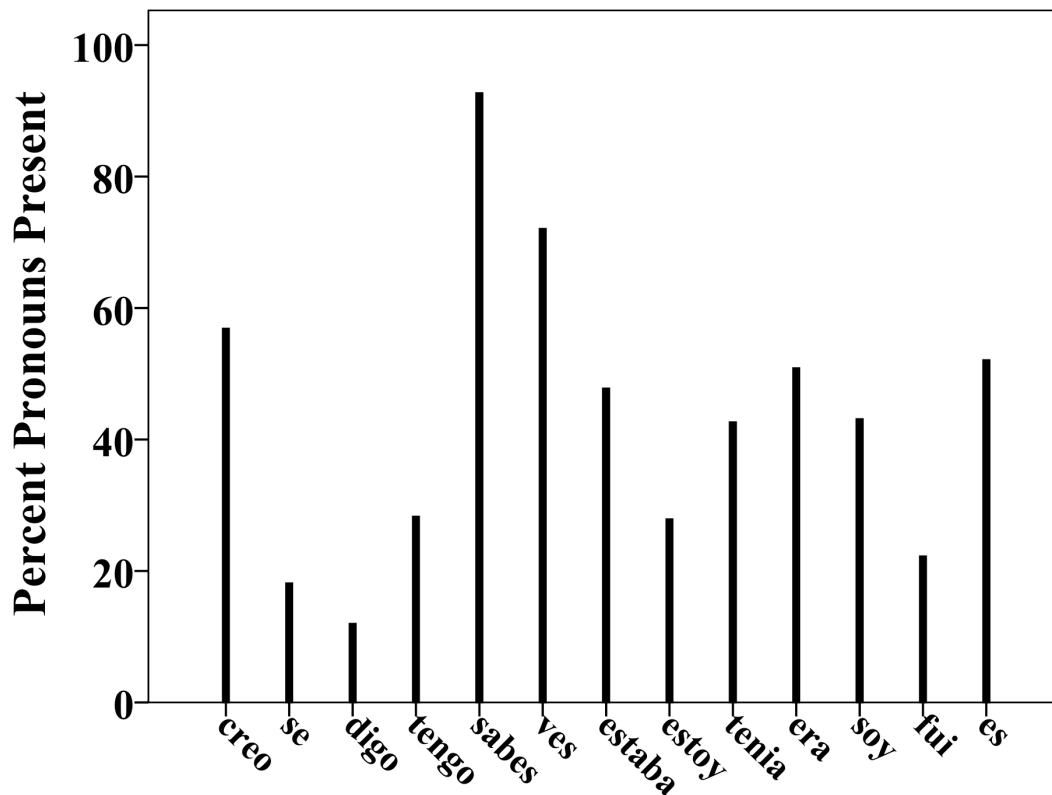


On its face, this result indicates that lexical frequency conditions pronoun use when it is treated as a discrete variable. But there are two caveats to such a conclusion. First, we have not yet examined whether the other known predictors of SPP use account for this

difference between high and low frequency forms; this is done in section 3.3. Second, given the evidence of Figures 2 and 3, where the relationship between frequency and SPP use resists monotonic description, it is important to test the coherence of the notion that lexical frequency has a singular, uniform effect on the way pronouns are used.

A close examination of the most frequently occurring verb forms tends to belie this conclusion. If it is true that higher frequency verbs tend to attract more subject pronouns, we should expect to observe relatively high SPP rates among verbs at the extreme high end of the frequency range. Furthermore, we might expect some convergence at this end of the range, because random variability should wash out with high Ns. But this is not what the data show. Rather, the highest frequency forms show great divergence in SPP rates. This is illustrated in Figure 5, showing the fifteen most frequent forms in the corpus. Contrary to any expectation of consistent favoring of pronoun occurrence, these SPP rates range across most of the spectrum of possible values, from a low of 12% pronoun usage with *digo* ‘I say’ (raw frequency 117) to a high of 92% for *sabes* ‘you know’ (raw frequency 82). Hence it is not surprising that a one-way ANOVA reveals significant differences in SPP rates among these fifteen forms ( $F = 25.1, p < .001$ ).

Figure 5. Percent SPP presence among highest frequency forms (from most to least freq).



Thus far, the evidence for a frequency effect is inconsistent and contradictory: some measures indicate more pronouns with more frequent verbs, one suggests the opposite, and the highest frequency forms are highly variable. This is very different from the highly consistent effects of the other linguistic constraints that we show in the next section. Therefore, frequency does not appear to be a simple, orthogonal addition to the set of predictors of SPP use. Rather, it turns out that frequency systematically interacts with the other linguistic constraints on SPP occurrence.

### 3.2. OTHER linguistic constraints on spp use.

*Morphological Regularity* – Verb forms with regular morphology occur with pronouns significantly more often than do irregular forms. Thirty-seven percent of regular forms occur with a pronoun, while only 30 percent of irregular forms do (see Table 3).

	N Verbs	% Overt Pronouns
Irregular forms	3,170	30
Regular forms	1,682	37
t(3,584) = 4.557, p<.001		

*Person and Number* – Rates of pronoun use differ significantly across various person and number combinations. At 49 percent, the highest rate of pronoun use is found among second-singular forms. This is followed by third and the first-singular forms, with 39 and 37 percent overt pronouns. The plural forms lag substantially, with rates of 18 percent in the first-plural and 13 percent in the second-plural (Table 4).

	N Verbs	% Overt Pronouns
1 <sup>st</sup> singular	2,455	37
2 <sup>nd</sup> singular	846	48
3 <sup>rd</sup> singular	702	39
1 <sup>st</sup> plural	305	18
3 <sup>rd</sup> plural	605	13
F(4, 4908) = 36.120 p<.001		

*Tense-Mood-Aspect* – Significant differences in pronoun rates also emerge among TMA combinations<sup>7</sup>. At 43 percent, imperfect indicative forms occur with SPPs most frequently. Unsurprisingly, imperative verbs almost never occur with a pronoun (1%). The compound tenses with a present indicative inflected form (the periphrastic future – e.g., *voy a comer* – at 36%, and the perfect indicative – e.g., *él ha dicho* – at 33%) have SPP rates very close to the simple present (36%). The other inflected TMA combinations are lower, with the present subjunctive (*hables*) at 31% and preterite indicative (*hablaste*) at 29% (see Table 5).

	N Verbs	% Overt Pronouns
Present Indicative	2,695	36
Preterite Indicative	877	29
Perfect Indicative	176	33
Imperfect Indicative	708	43
Periphrastic Future	140	36
Present Subjunctive	96	31
Imperative	110	1
Perfect Indicative	176	33
F(6, 4795) = 8.9, p < .001		

*Semantic Content* – Mental activity verbs occur with pronouns at a higher rate (45%) than stative verbs (36%), which themselves have a higher SPP rate than external activity verbs (31%; see Table 6)<sup>8</sup>.

	N Verbs	% Overt Pronouns
Mental Activity	840	45
Stative Verb	1,438	36
External Activity Verb	2,601	31
F(2, 4878) = 27.798, p < .001		

*Switch Reference* – As found in all prior studies, verbs that represent a switch in referent have a significantly higher SPP rate (see Table 7).

	N Verbs	% Overt Pronouns
Switch in Referent	2,653	40
Not a Switch in Referent	2,233	29
t = 8.1 p < .001		

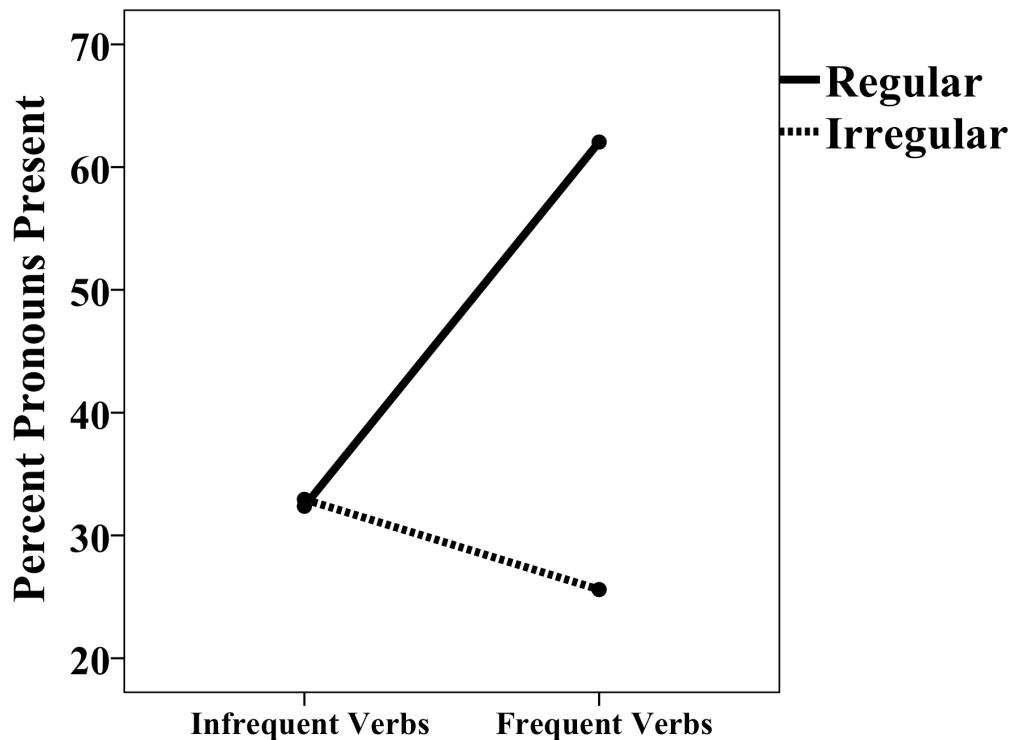
Each of these linguistic constraints has a significant effect on SPP use. The effects reported here are consistent with those found in previous research; indeed, they are very close in both direction and magnitude. Since our data set therefore presents a typical pattern of Spanish pronoun use, it is unlikely that any frequency effects

encountered here are due to something unusual about this corpus. But the systematicity of these constraints is strikingly at odds with the inconsistency of the frequency effects. Therefore, we now turn to an exploration of that inconsistency, by examining the interactions between these constraints and lexical frequency.

**3.3. LEXCIAL frequency – interaction with other constraints.** Is frequency orthogonal to and independent of the other factors that condition pronoun use, or does it interact with them? Given the relatively weak correlations between our continuous frequency measures and SPP rates, we pursue this question using the *discrete frequency* variable, by comparing the effects of the core constraints on the two frequency groups of verbs – frequent and infrequent. If frequency is independent and orthogonal, the main effects of the other constraints should be consistently observed in both frequent and infrequent verb forms.

As an example, recall the main effect of morphological regularity: regular forms occur with more subject pronouns (37%) than irregular forms (30%). Absent interaction, this relationship should be comparable in both low and high frequency verbs. As Figure 6 shows below, this is not the case. Instead, frequent and infrequent forms behave very differently with respect to morphological regularity. Among infrequent forms, there is no significant difference between regular and irregular verbs; in fact, their SPP rates are nearly identical – 32% for regulars and 33% for irregulars ( $F = .1, p = .74$ ). This contrasts starkly with the frequent forms, where regularity is robustly significant ( $F = 102.1, p < .001$ ), and the 62% occurrence rate for regular forms is more than double that of irregulars, 26%.

Figure 6. Morphological regularity – frequent vs. infrequent forms

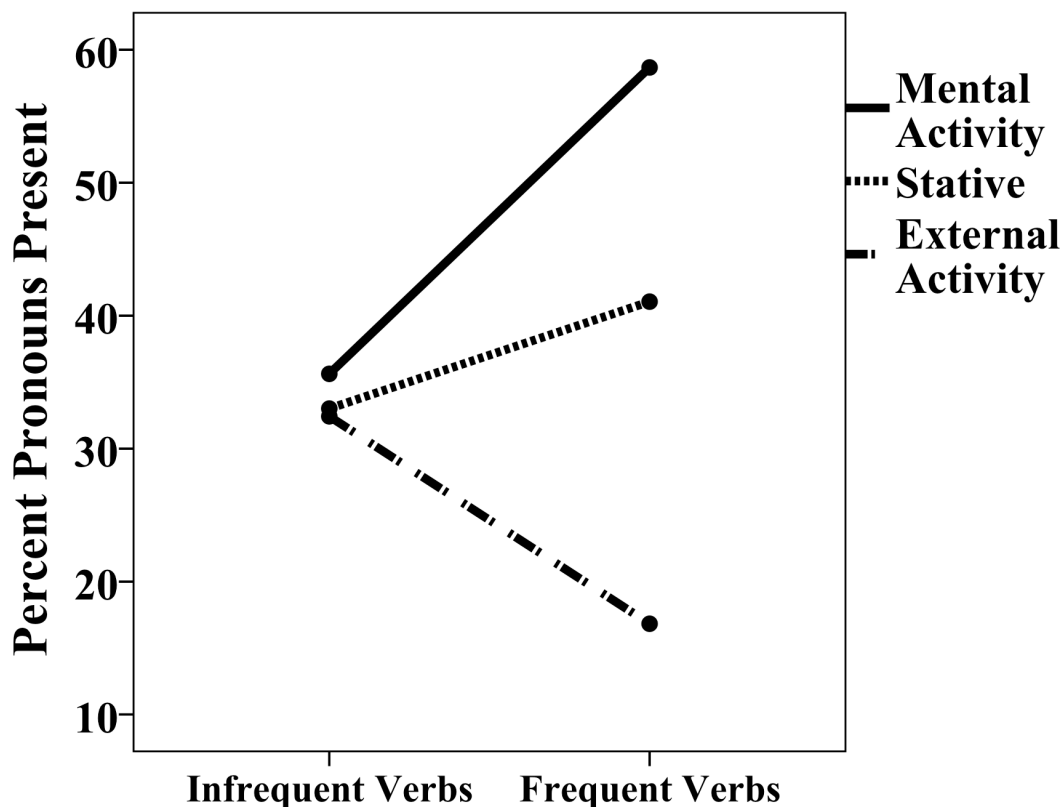


Clearly, frequency and morphological regularity interact. Moreover, these results imply that the main effect of morphological regularity observed above is due entirely to the frequent forms. And most notably, we observe that the effect of frequency is not uniform. The apparent main effect of *discrete frequency* above was that frequent forms attract more pronouns than infrequent forms, but here we see that this is not true of all verbs. Rather, in the irregular forms, the frequent words attract LESS pronoun usage – 26%, vs. 33% in the infrequent irregulars. The pronoun-favoring effect of high frequency is found only in the regular forms, which go up from from 32% in infrequent cases to 62% in the frequent forms.

Similar results emerge for each of the other core conditioning factors. Consider semantic content: the main effect above was that mental activity verbs occur with pronouns the most, followed by stative and external activity verbs, with SPP rates of 45%, 36%, and 31%, respectively. But when frequent and infrequent forms are considered separately, we see that the effect of semantic content is again restricted to the frequent forms (Figure 7). In infrequently occurring verbs, there is no significant

difference in SPP rates among the three semantic content classes ( $F = .94, p = .38$ ). But among the frequent forms, robust and significant differences appear ( $F = 51.2, p < .001$ ). Furthermore, high lexical frequency again has a non-uniform effect: comparing infrequent to frequent forms within each class, mental activity and stative verbs show increases in SPP rates with higher frequency – from 36% to 59% and 33% to 41% respectively – while in external activity verbs, higher frequency is associated with a decrease in SPP rate – from 32% down to 17%.

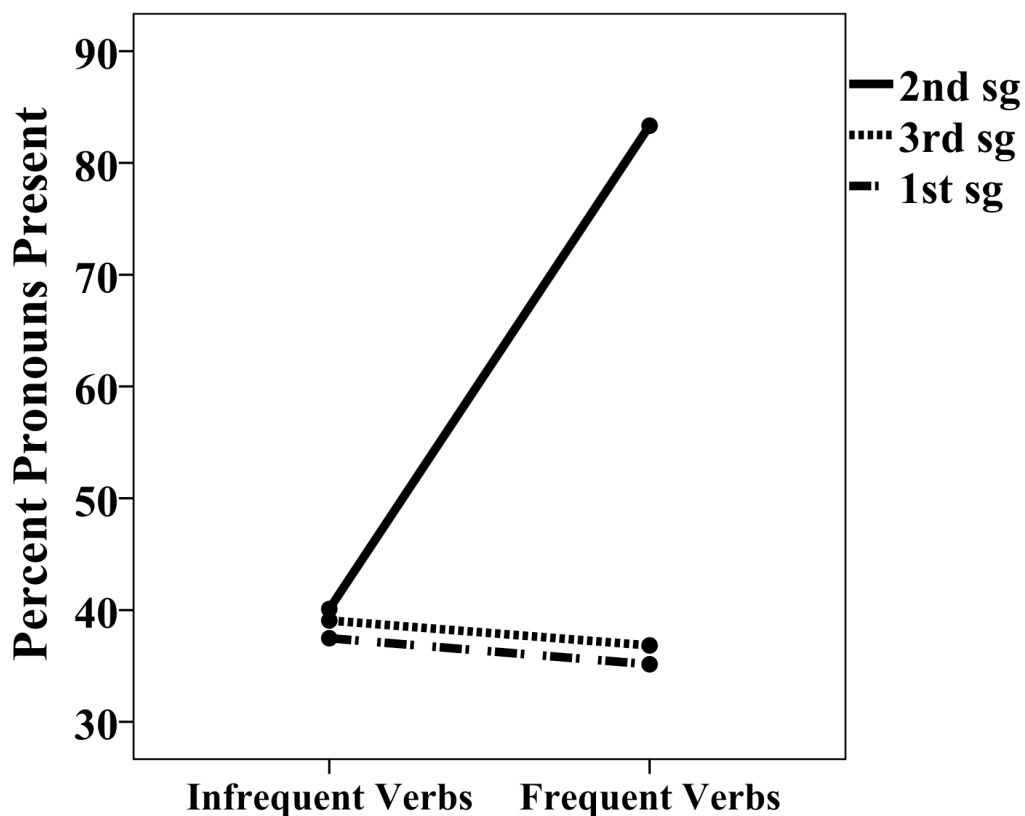
Figure 7. Semantic content – frequent vs. infrequent forms.



This pattern of interaction also appears in Figure 8 for the person/number variable<sup>9</sup>. Here again significant effects of the conditioning factor are limited to frequent forms. Among infrequent forms, pronoun use is very similar (and not significantly different,  $F = .7, p = .47$ ) across person and number combinations, but for frequent forms, they diverge significantly ( $F = 68.9, p < .001$ ). And again we observe that SPP rates do

not uniformly increase with frequency: higher lexical frequency is associated with a slight decrease in pronoun use among first and third singular verbs – from 37% to 35%, and 39% to 37% – but with a substantial increase in second singular forms – from 40% to 83%.

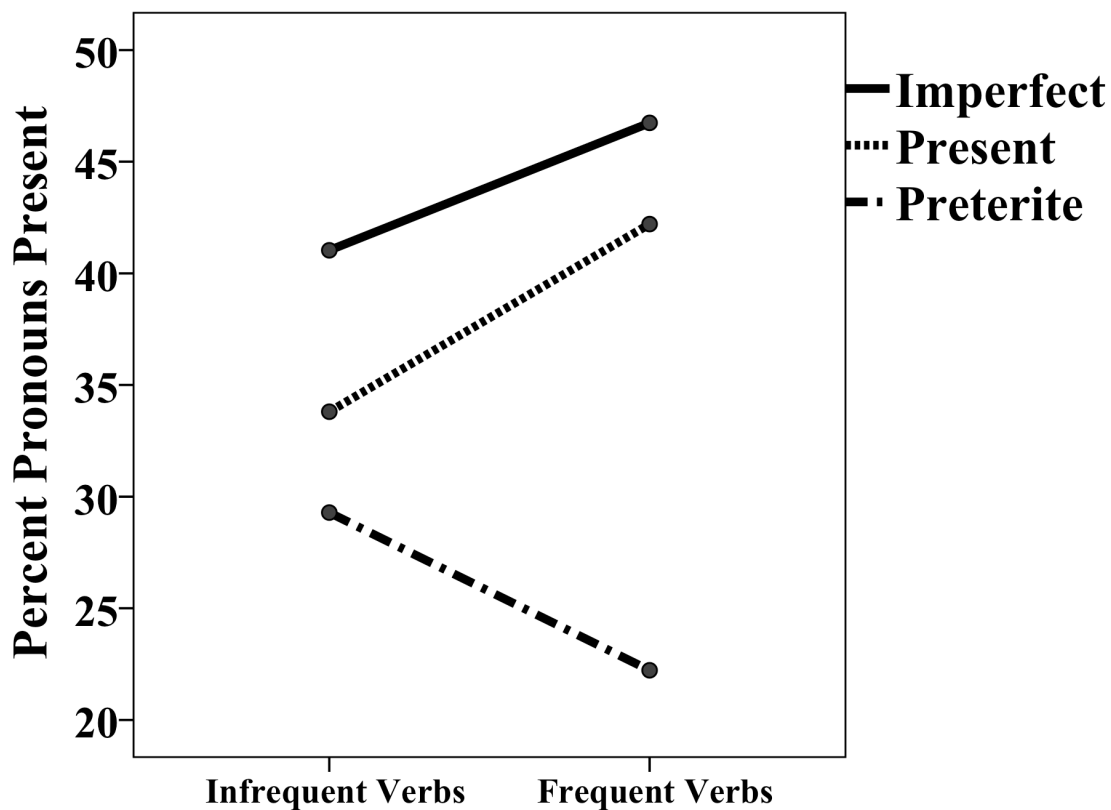
Figure 8. Person/number – frequent vs. infrequent forms.



Results for tense-mood-aspect and switch reference show the same general tendency of stronger effects among the more frequent verbs, but with one difference (see Figure 9). Both of these constraints show appreciably stronger effects among frequent forms, but the smaller effects among infrequent forms are still significant. First consider TMA<sup>10</sup>. Among infrequent forms, there are significant differences in pronoun rates for the imperfect (41%), present (34%), and preterite (29%) indicatives ( $F=9.9$ ,  $p<.001$ ). Among the frequent forms, we observe the same relative order, but the magnitude of the effect is much larger: 47% of frequent imperfect forms occur with a pronoun vs. only

22% of frequent preterit forms ( $F = 5.3, p < .001$ ). This range of 25% difference between the highest and lowest values is more than twice the range in SPP rates found for infrequent forms. And again, high frequency does not have a uniform effect. Rather, when compared to their infrequent counterparts, some frequent forms have higher SPP rates (here, the imperfect and present forms) while others have lower rates (the preterites).

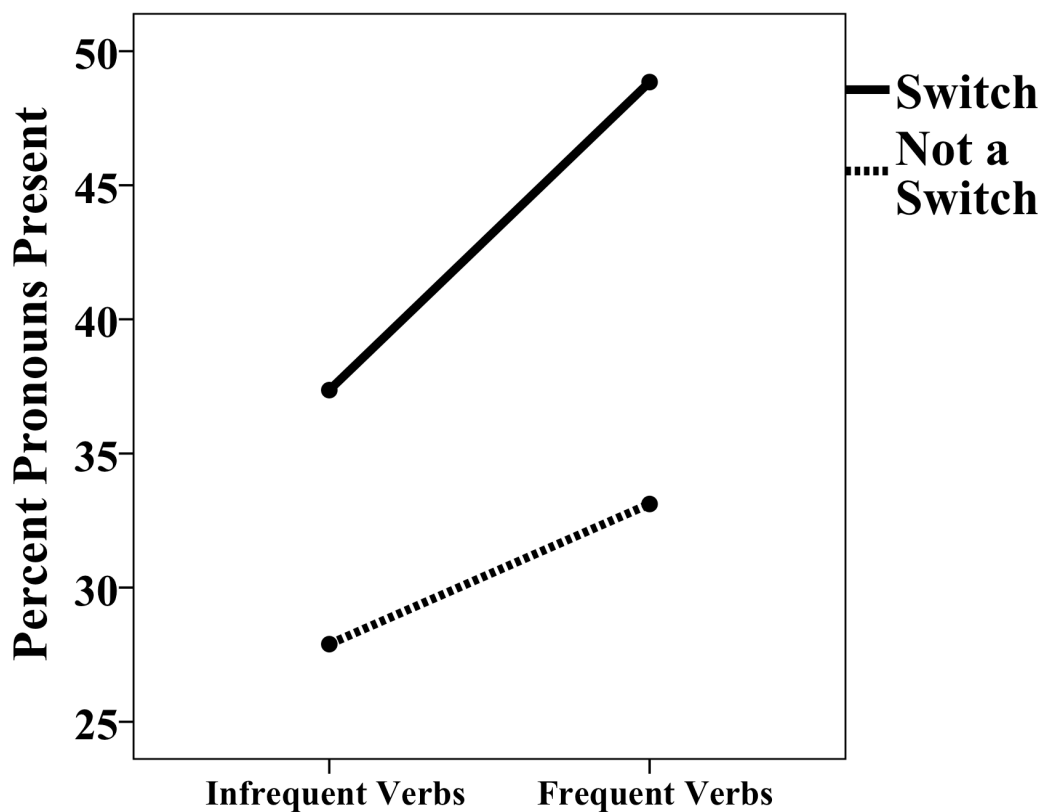
Figure 9. TMA – Frequent vs. infrequent forms.



Results for switch reference, shown in Figure 10, are similar to those for TMA in that significant differences are observed among both infrequent and frequent forms – switches in referent are always associated with significantly higher pronoun usage than unswitched references – but the magnitude of this effect is appreciably greater in high frequency forms (the range of values almost doubles, from 9% to 16%). An additional aspect of the results for switch reference that is unique among the preceding analyses is

that SPP rates for both frequent groups are higher than those of their infrequent counterparts. This is the only analysis in which high frequency appears to uniformly favor increased SPP use.

Figure 10. Switch reference – frequent vs. infrequent forms.



The results of the preceding analyses are summarized in Table 8, which also lists significance figures for tests of interactions between discrete frequency and each of the conditioning factors.

Conditioning factor	Main effect on infrequent forms	Main effect on frequent forms	Interaction with Frequency
<i>Morph Regularity</i>	NO p = .73	YES p < .001	YES p < .001
<i>Semantic Content</i>	NO p = .38	YES p < .001	YES p < .001
<i>Person &amp; Number</i>	NO p = .47	YES p < .001	YES p < .001
<i>TMA</i>	YES p < .001	YES p < .006	Near Sig. p < .077
<i>Switch Reference</i>	YES p < .001	YES p < .001	Near Sig. p < .057

The generalization that these results suggest is that lexical frequency does indeed play a crucial role in predicting SPP use, but not a monotonic or independent one. Rather, frequency operates by affecting the behavior of the core conditioning factors. Specifically, high frequency either activates or amplifies the predictive power of other constraints. For factors that are activated via interaction with frequency, their effects only appear among high frequency forms. This describes morphological regularity, semantic content, and person/number; these factors are entirely non-predictive among infrequent forms. The factors that are amplified, with stronger effects in high frequency forms, are TMA and switch reference. Figures 11 and 12 summarize these results.

Figure 11. Activation via Interaction

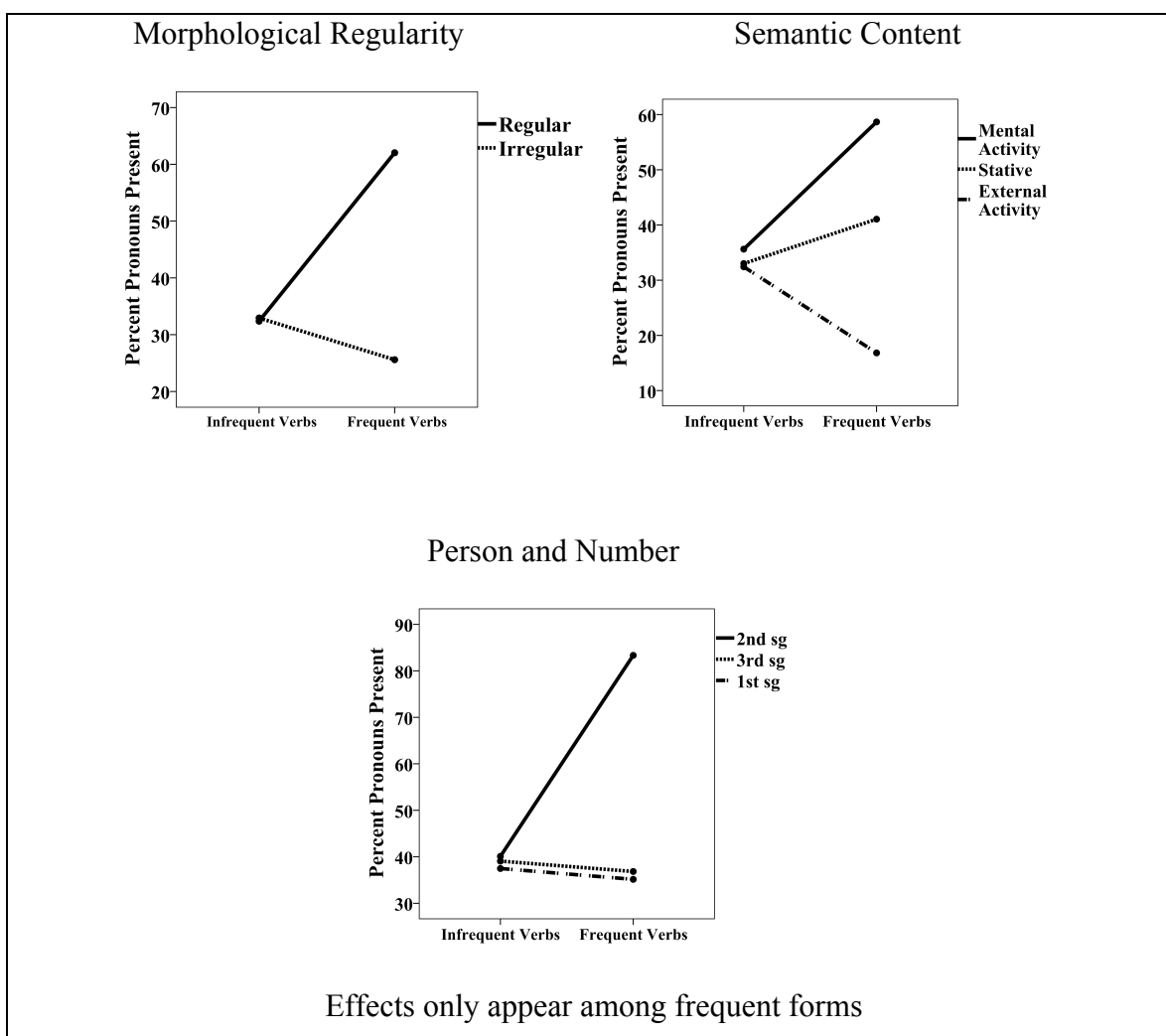
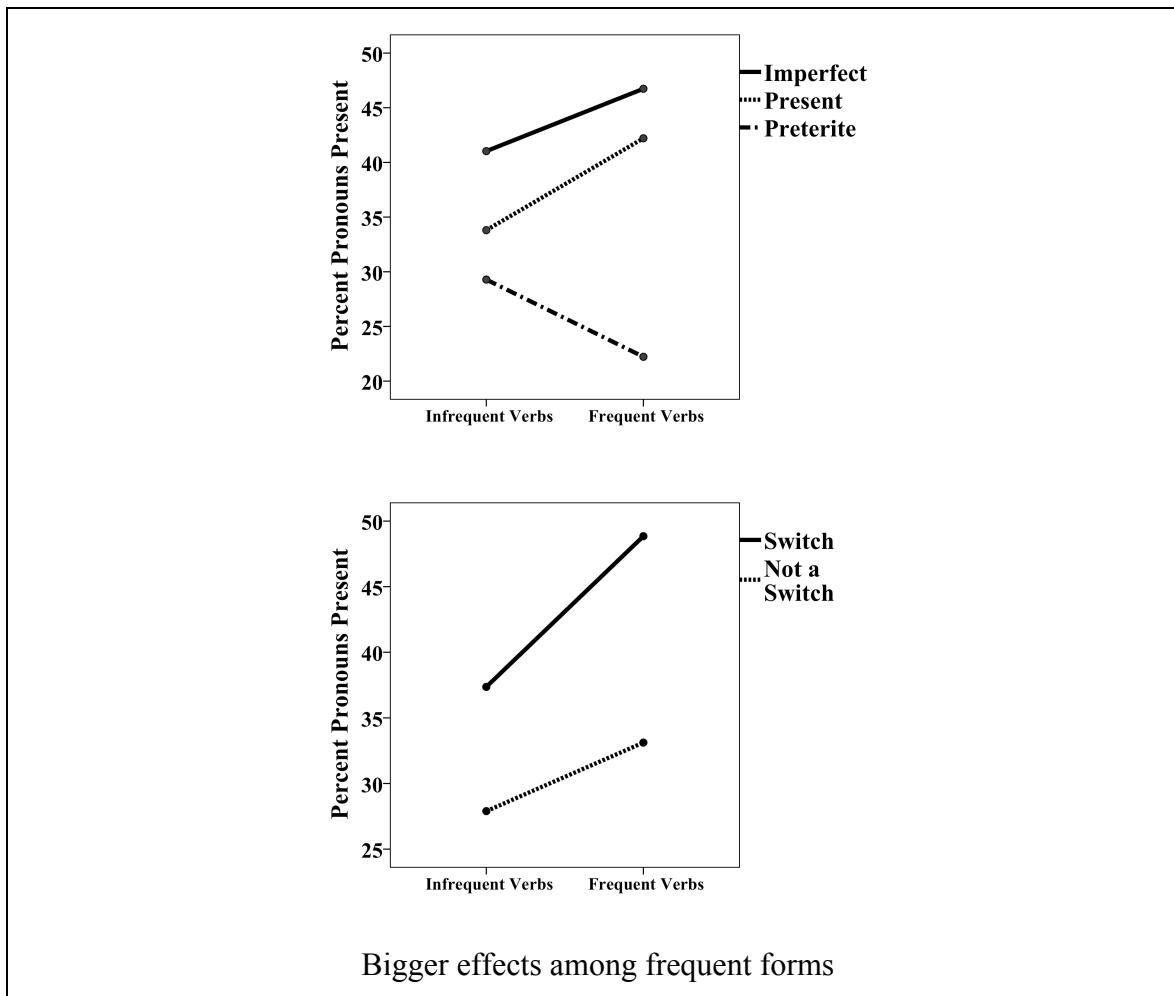


Figure 12. Amplification via Interaction

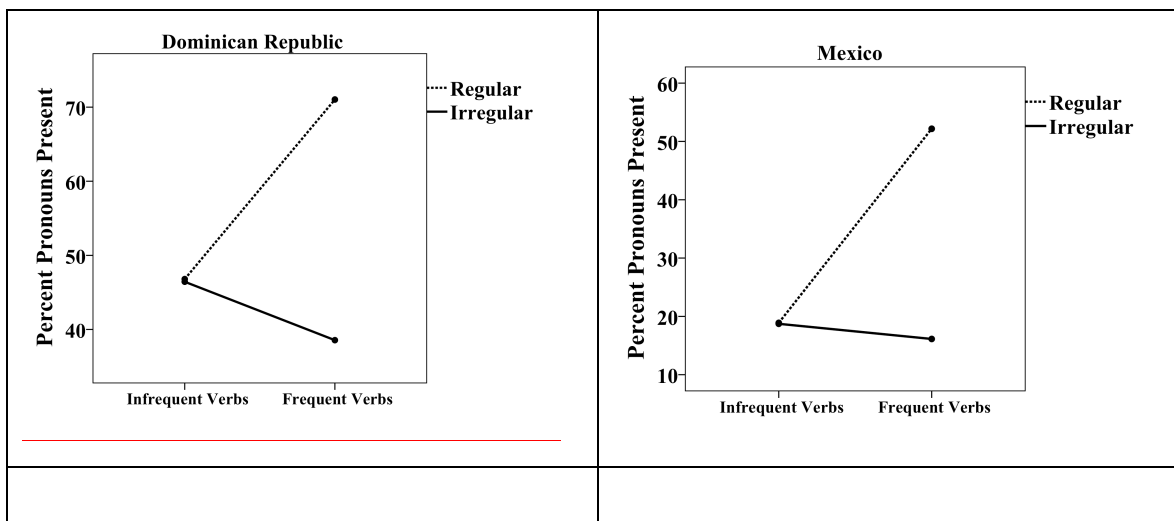


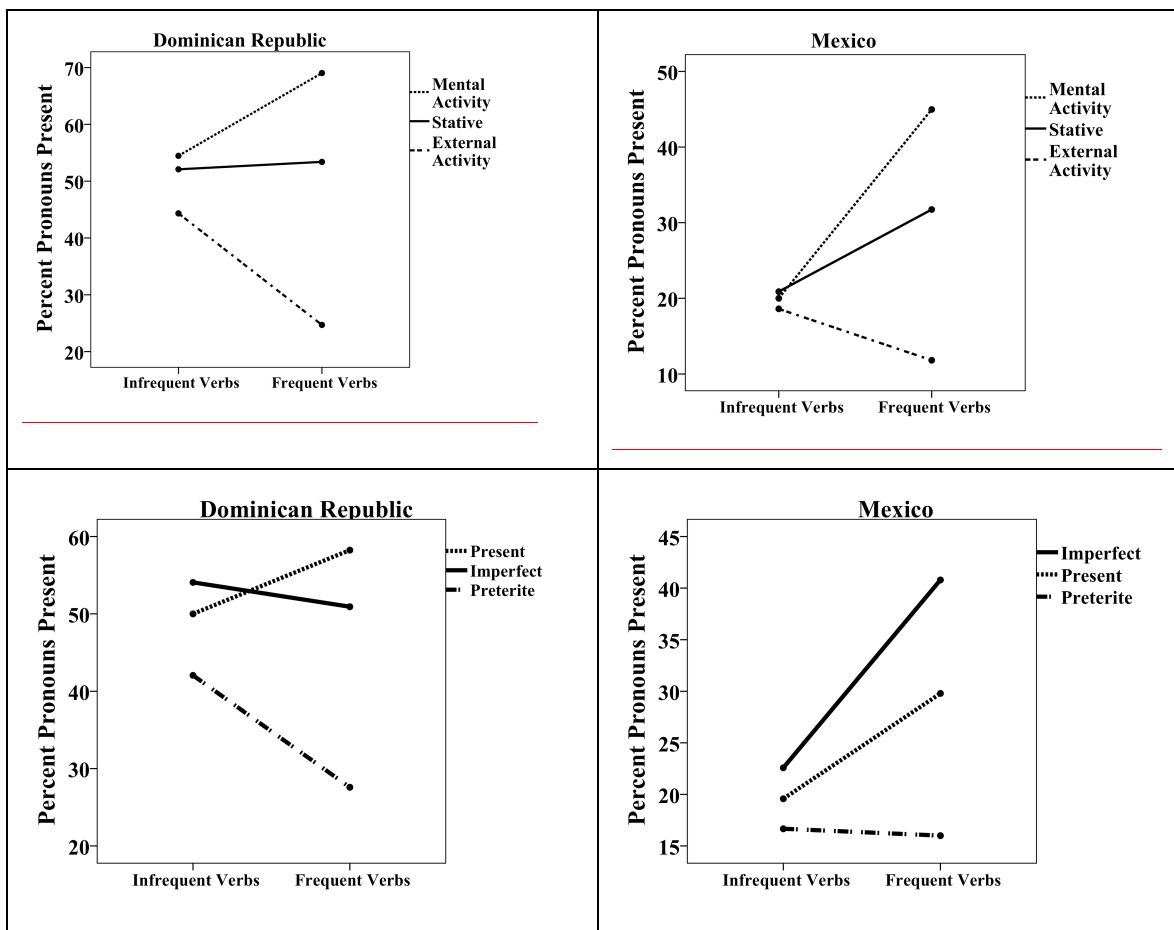
In light of these results, we conclude that high lexical frequency enables or potentiates other constraints on pronoun use. Indeed, each of the core constraints is weaker and less predictive of SPP use among infrequent forms and stronger and more predictive among frequent forms. We further conclude that frequency by itself does not have a uniform effect on pronoun use. Some categories within the core linguistic factors show increased pronoun rates at higher lexical frequency while others show reduced pronoun use. Thus it appears that lexical frequency does not make its own unique contribution to the probability that a form will occur with a pronoun or not. Regression results presented below confirm that this is the case.

**3.4. CONSISTENCY across dialects.** The systematic nature of the potentiation effect of frequency is further confirmed when we look separately at the two nationality groups in our sample. Both Dominicans and Mexicans show nearly identical patterns of interaction between frequency and other effects, despite the fact that their overall rates of pronoun use are dramatically different. Variationist research (Cameron 1993, Otheguy et al. 2007, Travis 2007) demonstrates that constraint effects provide a more penetrating characterization of grammatical structure than do overall rates of variable occurrence; hence the parallel behavior of these nationality groups indicates the systematicity of this finding.

Dominican speakers use pronouns at a significantly higher rate than Mexicans, 49% vs. 22% ( $t = 20.5$ ,  $p < .001$ ). But the constraint effects we have identified are consistently observed to be prominent in frequent verbs and weak or nonexistent in infrequent verbs, for both speaker groups, with only minor differences of detail. This is illustrated below for the morphological regularity, semantic content, and TMA of the verb.

*Figure 13.* Frequency effects by country





Similar trends emerge when speakers are separated according to their time of residence in New York City. The same frequency effects are found in the data for both the long-established New Yorkers and the recent arrivals from Latin America. These results are strong confirmation of the generality of the pattern: frequency affects pronoun use in a nearly identical way for subgroups in the sample that differ with respect to regional background and time in NYC, independently of their overall rates of pronoun use.

**3.5. MULTIVARIATE results.** If it is indeed true that lexical frequency does not independently affect the likelihood that a verb will occur with a pronoun or not, but instead only plays a role in pronoun use via interactions with other core constraints, then it should not be a significant predictor in a multivariate analysis that tests a model of SPP use which includes all relevant predictors simultaneously. Thus in a logistic regression

that includes discrete frequency alongside all the core linguistic factors, discrete frequency should not be a significant predictor of pronoun presence on its own. This is precisely what we find in Table 9, which shows the results of a logistic regression in which all the linguistic factors compete as predictors of the dependent variable, the presence or absence of pronouns. The analysis shows which constraints make a significant contribution to fitting the statistical model, and also shows the relative strength of the constraints, by the Wald value. The higher the Wald value for a given constraint, the stronger its role in conditioning pronoun use. The results show that each core linguistic factor is a significant predictor of pronoun occurrence, but frequency is not – it has no significant main effect. Nevertheless, frequency has a significant or nearly significant interaction with each core constraint.

Rank	Linguistic Constraint	Wald Value	Freq Interaction Terms
1	Person & Number**	144.2	F = 50.5**
2	Switch Reference**	76.9	F = 3.6 p <.057
3	Morphological Regularity**	36.5	F = 124.4**
4	TMA**	10.6	F = 2.6, p <.077 (near sig)
5	Semantic Content**	9.8	F = 33.8**
	Discrete Frequency (non-sig)	.157 (p = .692)	
**p <.001			

**4. DISCUSSION.** Our results show that lexical frequency does not have an independent effect on Spanish subject personal pronoun use in our data. Two continuous treatments of lexical frequency yielded weak and contradictory (although statistically significant) bivariate correlations with rates of pronoun occurrence: raw frequency ( $r = .057$ ,  $p <.001$ ) showed pronoun use increasing with frequency, but log frequency ( $r = -.081$ ,  $p <.001$ ) indicated the opposite. When frequency is treated as a discrete variable, there was a significant main effect – frequent forms have a higher rate of pronoun use – but this effect was not consistently observed among the most frequent verbs, and it disappears in a multivariate analysis that includes the other linguistic constraints. But despite lacking an independent main effect, frequency is not irrelevant to this variable; on the contrary, it has a strong and systematic, but complex effect, through interaction with all the other constraints.

This result is surprising on several accounts. First, it is inconsistent with typical multivariate models of linguistic variation. Such models ordinarily envision multiple constraining factors that, at least in their primary mode of operation, are independent and orthogonal and make separate probabilistic contributions to the outcome. This is the concept that underlies previous work on variable SPP occurrence in Spanish, which motivated the constraints we investigated. Interactions between predictors are commonly encountered among social factors, but they are not typically expected among linguistic constraints. Neither are they well understood when they occur, nor are they incorporated into theoretical models of optionality. But in the present data, we find systematic interaction between lexical frequency and all other linguistic constraints examined.

Second, the frequency effects reported are at odds with the simple effects of frequency that have been observed, or proposed on theoretical grounds, in much previous work, as discussed above. In phonology, Bybee argues for a monotonic frequency effect in phonological lenition: frequent repetition should favor phonological reduction, so high frequency items should lead synchronically in variation and diachronically in change. In psycholinguistic studies, high frequency monotonically favors quicker access and recognition, more favorable acceptability judgments, et cetera. In work on syntactic variability, higher frequency has been found to monotonically favor omission of complementizers and relativizers (cf. Jaeger 2006). But in variable SPP use in Spanish, higher lexical frequency of the verb is associated with reduced rates of subject pronoun use for some subsets of the data (e.g., morphologically irregular forms, external activity verbs), but with higher rates in others (e.g., morphologically regular forms, mental activity verbs). This makes simple and unidirectional accounts of frequency effects problematic, and leaves significant theoretical gaps in explaining how frequency works.

Third, these results raise questions about the status of the other linguistic constraints on pronoun occurrence. Three constraints considered here – morphological regularity, semantic content, and person/number – do not, in fact, operate across the entire lexicon; rather, they are significant only for the verbs in our high frequency category. For the remainder, which constitutes a huge majority of all the types in our data, and three-quarters of all tokens, they have no effect. And the other two – TMA and switch reference – are notably weaker in infrequent forms. So what is the theoretical

status of such constraints if they interact so dramatically with lexical frequency? If these factors capture generalizations or rules or constraints about when pronouns should be expressed, why are they contingent on frequency?

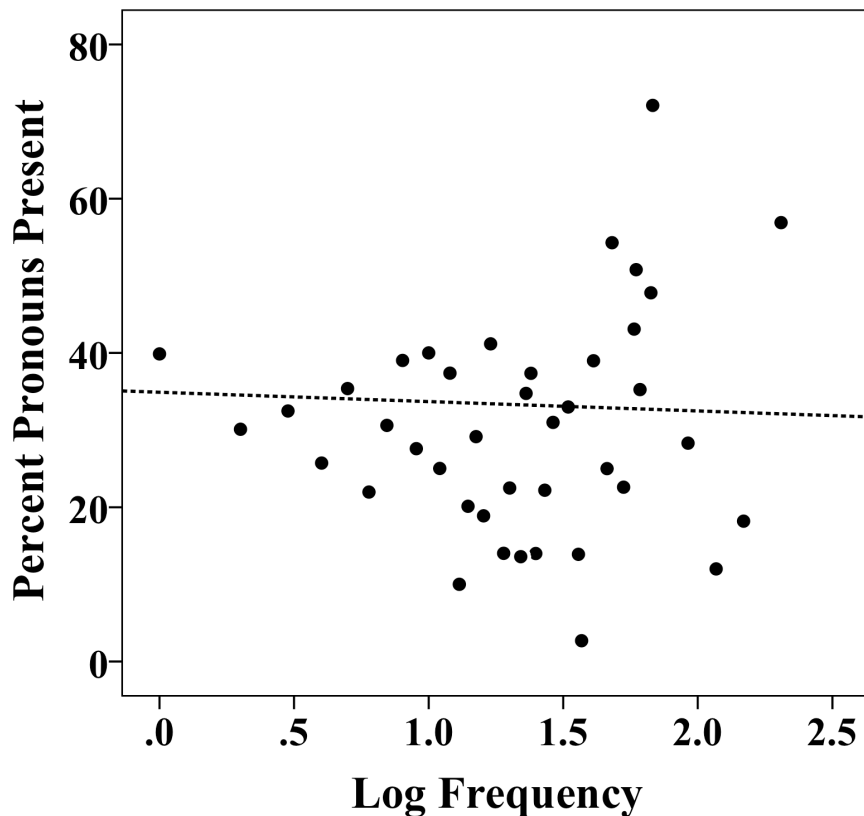
**4.1. FREQUENCY threshold.** The frequency effect observed here is not what on many grounds we might have expected it to be. It is not another independent predictor, it is not a modulation of the patterns associated with other constraints on pronoun use, it does not have a monotonic effect. But it is strong and systematic, if unexpected: all other constraints on pronoun use are weak or nonexistent among low frequency items, but strong and significant among high frequency items. So what is going on? We argue that higher lexical frequency acts as a potentiator or enabler for the other constraints – they are activated or amplified by high frequency. High frequency verb forms are the ones that speakers have more experience with and hear more often from others. Such forms can be expected to provide language users with richer and more reliable evidence about their contexts of occurrence, and their probable collocations. Low frequency forms, by contrast, provide little evidence about any distinctive properties they might have. We propose, therefore, that frequency affects linguistic variation not directly, but indirectly, as a limitation on the sufficiency of evidence. The frequency threshold is the level at which a speaker has enough evidence to formulate reliable estimates of distinctive lexical properties.

Consider how a lexical effect on pronoun occurrence must be acquired and remembered. An association between a specific verb form and the expression or non-expression of its subject pronoun is an item of collocational information; to accumulate such information a language learner must take note of the distribution of overt and null subjects and remember this pattern for use in production. This task is an order of magnitude more complex than just identifying the verb: the subject-verb relationship is a structural one, so identifying it requires a full parse of the utterance, rather than merely storing adjacency information. And not every token of a particular verb provides relevant information: those with full NP subjects do not illuminate the issue of pro-drop, conjoined VPs may lack overt subjects due to other syntactic operations, et cetera.

Consequently the learning task is subject to significant practical limitations. To learn the typical collocations of a verb, a speaker needs a sufficient sample of relevant constructions, and more frequently occurring items provide better samples. Low frequency items may simply fail to occur often enough for a speaker to infer a reliable estimate of their collocations with pronominal subjects. From this perspective, then, frequency acts as a gatekeeper. Below some frequency threshold, items are too rare to formulate rich representations that include collocational information. Above the threshold, language users have sufficient information about each lexical item to individuate them with respect to collocations and syntactic operations.

**4.2. SPECIFYING the threshold.** If this interpretation is valid, it is worth inquiring what this threshold value is. Our criterion for dividing frequent from infrequent forms was set somewhat arbitrarily at 1% of all verb forms; is this a reasonable value? This question can be approached empirically, looking again at the log frequency distribution of SPP rates (Figure 3 repeated here as Figure 14).

*Figure 14.* Log frequency and percent SPPs present



At the high frequency (right) end of the scale, we see the disparity of rates among high frequency forms noted above. Impressionistically, this dispersion is most pronounced among forms above a log frequency value of 1.5, which corresponds to a raw frequency value of about 33, or two-thirds of 1% of the data. In the analyses above we used threshold value of 49 (log value of 1.72, 1% of the data) for our ‘high frequency’ category. We can test the reasonableness of this criterion, and the validity of the findings, by redoing the analyses using log 1.5 as the cut-off figure. The results of such a reanalysis are presented in Table 10 (compare our original results above in Table 8.)

Conditioning factor	Main effect on infrequent forms	Main effect on frequent forms	Interaction with Frequency
<i>Morph Regularity</i>	NO p = .78	YES p < .001	YES p < .001
<i>Semantic Content</i>	NO p = .44	YES p < .001	YES p < .001
<i>Person &amp; Number</i>	NO p = .64	YES p < .001	YES p < .001
<i>TMA</i>	YES p < .001	YES p < .001	<i>Non Sig. p &lt; .24</i>
<i>Switch Reference</i>	YES p < .001	YES p < .001	<i>Non Sig. p &lt; .25</i>

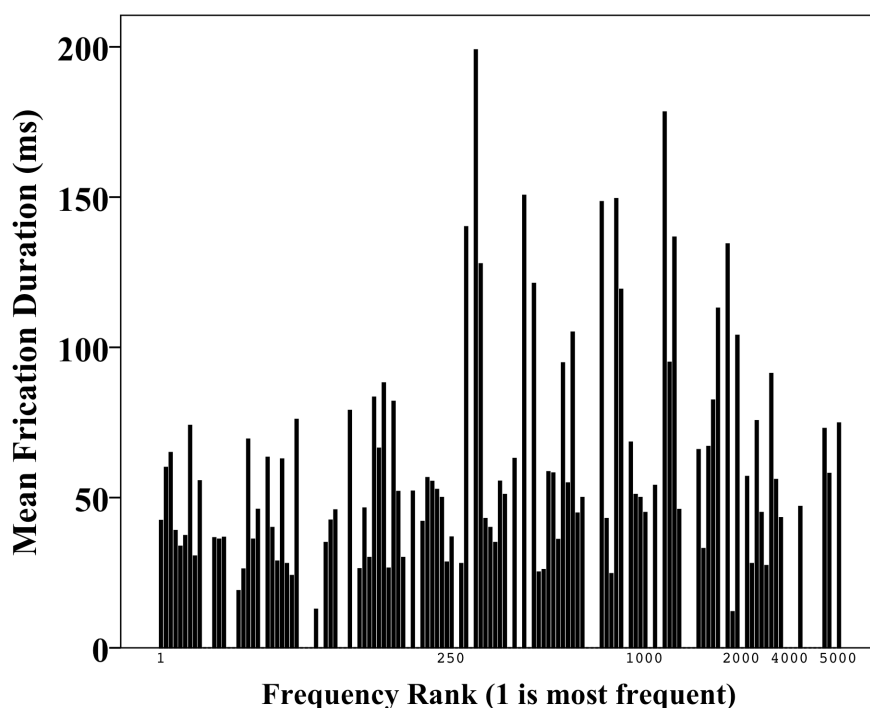
The reanalysis with a criterial value of log 1.5 yields results that are nearly identical to those originally obtained using log frequency 1.72. The presence or absence of significant main effects is exactly the same; the only notable difference between this and the previous analysis is that the interactions between frequency and TMA and Switch Reference are no longer near-significant. We suggest that this supports our model: the polarization of constraint effects by frequency category – stronger effects at higher frequencies – is indeed greatest at the highest frequency end of the scale. Lowering the criterial value brings into the ‘high frequency’ category some additional forms that show weaker effects of these two constraints, thus watering down the difference between high and low frequency forms, and reducing the significance of the interaction. As such, this is evidence against a lower criterial value.

Does the difficulty of identifying a precise threshold value imply that we should return to treating frequency as a continuous variable? We think not. As we have seen, the ‘long tail’ of low frequency items does not show consistent differentiation of pronoun use, neither by frequency, nor by most of the other constraints. We have further seen that

continuous treatments of frequency yield inconsistent results. Rather, frequency is best captured in these data by an inflection point, below which pronoun rates are not well differentiated, and above which all manner of factors come into play as predictors and differentiators of pronoun use. This inflection point is the threshold, and appears to lie in the vicinity of log 1.5-1.7 in these data.

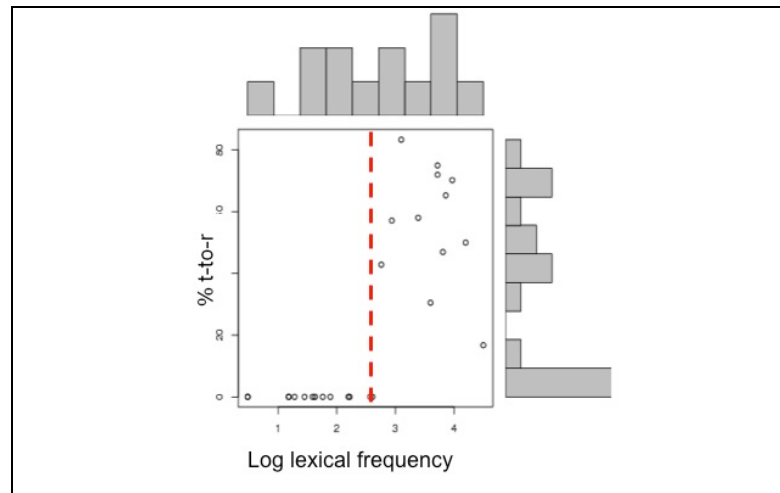
**4.3. FURTHER considerations.** The patterns we have found – a threshold effect of frequency, and disparate behavior among high frequency items arising from interaction – are evident in other work on frequency. For example, in work on frequency and phonetic gradience, Erker (2011) finds a threshold effect for s-lenition in Dominican Spanish, as shown in Figure 15. The 250 most frequent words in the data set have consistently short durations of coda /s/. Beyond this frequency cut-off, mean durations are significantly longer. But when frequency was treated continuously, it was not significant, and neither was there any significant differentiation of words by frequency within each frequency group.

*Figure 15.* Lexical frequency and /s/ lenition in Dominican Spanish.



A similar, dramatic threshold effect appears in Watson and Clark's (2010) work on 'rhotacization' of intervocalic /t/ in Liverpool (substitution of a tap in place of a fully articulated /t/). Lexical items with log frequencies below about 2.5 show no evidence whatsoever of this phonological phenomenon, but every word above that level shows some appreciable percentage of rhotacization (see Figure 16).

Figure 16. Threshold effect in t-to-tap in Liverpool (from Watson 2010)



All these results are consistent with the model we have suggested, in which there is a frequency threshold above which speakers have sufficient evidence to learn and reproduce lexically specific effects, but below which the paucity of evidence inhibits acquisition of special characteristics, including lexically individuated rates of variable phenomena, whether tapping, sibilant shortening, or pronoun occurrence. Indeed, the generalizations made by Bybee (2001) and Phillips (1986) about high frequency forms leading phonological variation and change are also consistent with this model: a lexical item must be encountered frequently in order to be accorded any special status.

It is also worth noting that our data are not unusual with respect to the nonuniform behavior among the highest frequency items seen in Figure 5. A case in point is Brown and Cacoullos (2002) work on /s/ lenition in Spanish. Despite the implicit prediction of Bybee's model that the highest frequency items should be most advanced in phonological reduction, Brown and Cacoullos find substantial divergence among the most frequent

words. For the coda /s/ data, among the six most frequent words in their sample, they find lenition rates ranging from a high of 80% for *mismo* ‘same’ (the 6<sup>th</sup> most common word) to a low of 11% for *este* ‘this’ (2<sup>nd</sup> most common word), with the most common word *estar* ‘be’ in between at 34%. In syllable-initial position, the same disparities occur. For example, highest frequency word *sí* ‘yes’ has a lenition rate of 29%, much lower than the 10<sup>th</sup> most frequent word *pasar* ‘pass’, 58%. The highest lenition rate of 84% is found in *nosotros* ‘we’, which occurs at only one-tenth the frequency of *sí*. Such disparities raise the possibility that frequency effects may also be non-orthogonal and non-monotonic in phonology, just as we find in our syntactic case (for a similar result in syntactic variation, see Cacoullos and Walker 2009). Subsequent investigations would do well to test for interaction between frequency and other constraints on variable processes.

Our general point, that high frequency items provide sufficient evidence for learning patterns, but low frequency items do not, is echoed in an observation by Jaeger (2006: 156), in his research on complementizer and relativizer deletion in English. He finds interaction between lexical frequency and other constraints on deletion, such that higher frequency magnifies the constraint effects. He suggests:

...the PRH [his hypothesis that higher predictability of items favors reduction/deletion] is compatible with weaker effects for low-frequency cues.... From a learner’s perspective low frequency embedding verbs and head nouns are less available cues. That is, even if a low frequency cue is rather reliable (i.e. if it contains a lot of information about the likelihood of an outcome event), it may still be too rare for learners to conclude that they have reliable information about it.

This is essentially what we observe in the Spanish pronoun case: weaker or absent constraint effects at lower frequencies. But this kind of interaction also renders problematic the status of the supposedly independent constraints. If they are weaker, or even disappear, at low frequencies, it may be inappropriate to think of them as abstract conditioning factors that govern pronoun use with general applicability across the

grammar and the lexicon. Instead, it may be more useful to treat the conditioning factors as emergent generalizations based on the behavior of individual lexical items, specifically those that occur at a high enough frequency to permit reliable inferences about collocations. This is consistent with a Bybee approach, and it would furthermore require lexical representations that are rich enough to store information about frequency and collocations, as postulated by usage-based models.

However, evidence in support of abstract generalizations ('rules') that are independent of lexical items can also be found in these results: the switch reference and TMA variables are predictive of SPP use for both frequent and infrequent words. Furthermore, the interaction of these two variables with frequency is weaker than the other interactions in our analysis. This may indicate a fundamental distinction between discourse-level constraints and constraints that are systemic to the verbs forms themselves. Just such a distinction has been emphasized in studies of L2 acquisition of Spanish (Geeslin and Gudmestad 2008 and forthcoming). While an individual verb form is invariant in its systemic characteristics – for example, *canto* 'I sing' is always first person singular, regular, and represents an external activity – its status for switch reference varies as a function of the discourse. Whether it represents a switch or a continuity of reference depends on preceding discourse. In this sense switch reference is orthogonal to lexical selection. While not entirely equivalent, variability in TMA is also discourse dependent, certainly more so than morphological regularity, semantic content, and person/number. To the extent that the tense and mood properties of individual verb forms vary as a function of the temporal aspects of a discourse, it may be worth considering switch reference and TMA as qualitatively different kinds of linguistic factors. From this perspective, one might speculate that among the possible linguistic constraints on pronoun use in Spanish, those that operate at (or closer to) the level of discourse are more likely to apply generally, to constitute rules that are not lexically individuated. By comparison, the predictive power of constraints that are systemic to individual forms will be restricted to high frequency lexical items. This suggests that existing analyses of SPP use may need to be reconsidered.

**4.4. IMPLICATIONS** for the study of lexical frequency and linguistic variation. The results we have presented have important implications for further quantitative study of both frequency effects and the other constraints that are routinely investigated as predictors of variability. On the frequency side, future investigations of frequency may need to examine whether it is better treated as continuous or discrete, whether it shows a threshold effect, whether it is monotonic, and whether it shows interaction with other constraints. Even a finding that frequency has no significant main effect may not be sufficient reason to neglect it, if it potentiates other constraints. Since our understanding – both empirical and theoretical – of frequency is still in its early stages, researchers are well advised to approach lexical frequency from as many angles as possible.

The investigation of other constraints on variable processes is also problematized by these findings. The factors considered here are robustly replicated in numerous studies, and were generally assumed to operate across the lexicon as independent predictors. Our results show these factors to have strong and systematic interactions with frequency, and to often disappear in the long tail at the low end of the frequency distribution. If linguistic constraints are potentiated by frequency, then a principal focus of any study that includes frequency should be interaction. Methodologically, it may be inadequate to simply add frequency to a set of other constraints in a regression analysis, for example, by adding a frequency factor group to a Varbrul analysis. If the linguistic constraints on a process are emergent generalizations, then attention must be paid to how ‘general’ they are, and where they fail, lexically speaking.

**5. CONCLUSION.** The results presented here show that lexical frequency effects on variable subject personal pronoun expression in Spanish interact systematically with morphosyntactic, semantic, and discursive constraints that are known to affect this variable: in all cases higher frequency amplifies other constraints. This is subject to a discontinuous threshold effect: above a certain frequency level, there is a significant expansion of differentiation among verb forms, and of the strength and significance of linguistic constraints on pronoun occurrence. We propose a model that accounts for these facts: frequency does not directly constrain syntactic or other linguistic operations; rather it constrains the evidence that speakers have about these operations. Rarer forms provide

insufficient or unreliable evidence about lexically specific aspects of variable processes – in this case, the co-occurrence of a verb with an overt subject pronoun. Frequency effects are thus mediated through acquisition. We suggest that this model generalizes to account for reported frequency effects on phonological variation and change.

These results raise methodological and theoretical questions. Methodologically, more careful treatments of frequency are indicated, using analyses that incorporate interaction terms, and partition the data. From the standpoint of theory, these results favor models that incorporate rich lexical representations containing information about frequency, like those proposed in usage-based and exemplar approaches to language. The relative instability of the other constraints on Spanish SPP expression – appearing among high frequency forms but diminishing to insignificance among low frequency forms – raises questions about their status as independent predictors. This pattern is consistent with such classes of verb forms constituting emergent generalizations, rather than primary categories.

The essence of lexical frequency, then, is not practice but experience. A speaker's knowledge of a word is underdeveloped until it has been frequently heard and used. In the words of Clarence Day, information is pretty thin stuff unless mixed with experience.

## REFERENCES

- Alba, Matthew C. 2008. Ratio frequency: Insights into usage effects on phonological structure from hiatus resolution in New Mexican Spanish. *Studies in Hispanic and Lusophone Linguistics* 1.2: 247-286.
- Amaral, Patrícia Matos and Scott Schwenter. 2005. Contrast and the (non-)occurrence of subject pronouns. *Selected proceedings of the 7th Hispanic Linguistics Symposium*, ed. by David Eddington, 116–27. Somerville, MA: Cascadilla.
- Avila-Jiménez, Bárbara I. 1995. A sociolinguistic analysis of a change in progress: pronominal overttness in Puerto Rican Spanish. *Cornell Working Papers in Linguistics* 13.25-47.
- Barrenechea, Ana María, and Alicia Alonso. 1977. Los pronombres personales sujetos en el español hablado en Buenos Aires. *Estudios sobre el español hablado en las principales ciudades de América*, ed. by Juan M. Lope Blanch, 333-49. México: Universidad Nacional Autónoma de México.
- Bayley, Robert, and Lucinda Pease-Alvarez. 1996. Null and expressed pronoun variation in Mexican-descent children's Spanish. *Sociolinguistic variation: data, theory, and analysis*, ed. by Jennifer Arnold, Renée Blake, and Brad Davidson, 85-99. Stanford, CA: CSLI.
- Bayley, Robert & Pease-Álvarez, Lucinda. 1997. Null pronoun variation in Mexican-descent children's narrative discourse. *Language Variation and Change*, 9, 349-371.
- Bayley, Robert. 2002a. The quantitative paradigm. In Chambers et al., 117–41.
- Bentivoglio, Paola. 1987. *Los sujetos pronominales de primera persona en el habla de Caracas*. Caracas: Universidad Central de Venezuela.
- Bentivoglio, Paola. 1993. Full NPs in spoken Spanish: A discourse profile. In William Ashby, Marianne Mithun, Giorgio Perissinotto & Eduardo Raposo (Eds.), *Linguistic perspectives on Romance languages. Selected papers from the 21<sup>st</sup> LSRL* (pp. 211-224). Amsterdam: John Benjamins.
- Brown, Esther L., and Rena Torres Cacoullos. 2002. ¿Qué le vamoh aher? Taking the syllable out of Spanish /s/ reduction. *University of Pennsylvania Working Papers in Linguistics (PWPL)* 8.3, Papers from NWAV 30, 17-31.

- Brown, Esther L. 2004. Reduction of syllable initial /s/ in the Spanish of New Mexico and Southern Colorado: A usage-based approach. Dissertation, University of New Mexico, Albuquerque.
- Bybee, J. and S. Thompson. 1997. Three frequency effects in syntax. BLS, 23rd annual Meetings, 378-88. Berkeley: Berkeley Linguistics Society.
- Bybee, Joan. 2000. *The Phonology of the lexicon*. Usage-based models of language. Edited by Suzanne Kemmer and Michael Barlow. CSLI Publications.
- Bybee, Joan. 2000. Lexicalization of sound change and alternating environments. Papers in Laboratory Phonology V. Acquisition and the Lexicon. Edited by Janet Pierrehumbert and Michael B. Broe. Cambridge University Press.
- Bybee, Joan. 2001. *Phonology and language use*. Cambridge University Press.
- Bybee, J. and Hopper, P. (eds.) 2001. Frequency and the emergence of linguistic structure. Amsterdam: Benjamins.
- Bybee, J. L. 2002 Cognitive processes in grammaticalization. In M. Tomasello, editor, *The New Psychology of Language*, volume II. New Jersey: Lawrence Erlbaum Associates Inc.
- Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*. 14. 261-290.
- Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In B. D. Joseph and J. Janda (eds.) *The Handbook of Historical Linguistics*. Oxford: Blackwell. 602-623.
- Bybee, Joan and Rena Torres Cacoullos. 2008. Phonological and grammatical variation in exemplar models. *Studies in Hispanic and Lusophone Linguistics* 1(2).399-413.
- Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge University Press.
- Torres Cacoullos, Rena, and James A. Walker. 2009. The present of the English future: Grammatical variation and collocations in discourse. *Language* 85: 321-354.
- Cameron, Richard. 1993. Ambiguous agreement, functional compensation, and non-specific *tú* in the Spanish of San Juan, Puerto Rico and Madrid, Spain. *Language Variation and Change*, 5, 305-334.

- Cameron, Richard. 1994. Switch reference, verb class and priming in a variable syntax. In Katharine Beals (ed.), *Papers from the 30<sup>th</sup> regional meeting of the Chicago Linguistics Society: Volume 2: The parasession on variation in linguistic theory* (pp. 27-45). Chicago Linguistics Society.
- Cameron, Richard. 1995. The scope and limits of switch reference as a constraint on pronominal subject expression. *Hispanic Linguistics* 6/7, 1-28.
- Cameron, Richard. 1996. A community-based test of a linguistic hypothesis. *Language in Society* 25.61–111.
- Cameron, Richard. & Flores-Ferrán, Nydia. 2004. Perseveration of subject expression across regional dialects of Spanish. *Spanish in Context, 1*, 41-65.
- Cedergren, Henrietta and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50.333–55.
- Comajoan, Llorenç. 2005. Continuity and episodic structure in Spanish subject reference. In J. Clancy Clements & Jiyoung Yoon (eds.), *Functional approaches to Spanish syntax: Lexical semantics, discourse and transitivity* (pp. 53-79). London: Palgrave.
- Davidson, Brad. 1996. Pragmatic weight and Spanish subject pronouns: The pragmatic and discourse uses of *tú* and *yo* in spoken Madrid Spanish. *Journal of Pragmatics*, 26(4), 543-565.
- Ellis, Nick C. 2002a. Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*. 24, 143-188
- Ellis, Nick C. 2002b. Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*. 24, 297-339
- Enríquez, Emilia V. 1984. El pronombre personal sujeto en la lengua española hablada en Madrid. Madrid: Consejo Superior de Investigaciones Científicas.
- Erker, Daniel. 2011. An acoustic sociolinguistic analysis of variable coda /s/ production in the Spanish of New York City. PhD Dissertation. New York University
- File-Muriel, Richard and Brown, Earl. 2011. The gradient nature of s-lenition in Caleño Spanish. *Language Variation and Change*, 23, 223-243

- Flores-Ferrán, Nydia. 2002. Subject personal pronouns in Spanish narratives of Puerto Ricans in NYC: A sociolinguistic Perspective. Munchen: Lincom Europa.
- Flores-Ferrán, Nydia. 2004. Spanish subject personal pronoun use in New York City Puerto Ricans: Can we rest the case of English contact? *Language Variation and Change*, 16, 49-73.
- Flores-Ferrán, Nydia. 2005. La expresión del pronombre personal sujeto en narrativas orales de puertorriqueños de Nueva Cork. In Luis Ortiz López & Manel Lacorte (Eds.), *Contactos y contextos lingüísticos: el español en los EEUU y en contacto con otras lenguas*, pp. 119-129. Madrid: Iberoamericana.
- Flores-Ferrán, Nydia. 2007. Los Mexicanos in New Jersey: Pronominal expression and ethnolinguistic aspects. In Jonathan Holmquist, Augusto Lorenzino & Lofti Sayahi (Eds.) *Selected proceedings of the Third Workshop on Spanish Sociolinguistics* (pp. 85-91). Somerville, MA: Cascadilla Proceedings Project.
- Flores-Ferrán, Nydia. & Toro, Jeannette. 2000. The persistence of dialect features under conditions of contact and leveling. *Southwest Journal of Linguistics*, 19(2), 31-42.
- Francis, W. N. and H. Kučera. 1979. A Standard corpus of present-day edited American English, for use with digital computers. Brown University. Providence, Rhode Island.
- Geeslin, Kimberly & Gudmestad, Aarnes. 2011. Using sociolinguistic analyses of discourse-level features to expand research on L2 variation in forms of Spanish subject expression. Selected Proceedings of the 2009 Second Language Research Forum, ed. Luke Plonsky and Maren Schierloh, 16-30. Somerville, MA: Cascadilla Proceedings Project.
- Geeslin, Kimberly & Guijarro-Fuentes, Pedro. 2006. The second language acquisition of variable structures in Spanish by Portuguese speakers. *Language Learning*, 56(1), 53-107.
- Gudmestad, Aarnes. 2008. Acquiring a variable structure: An interlanguage analysis of second-language mood use in Spanish. (Doctoral dissertation, Indiana University, 2008). *Dissertation Abstracts International*, 69, 8.
- Gudmestad, Aarnes & Geeslin, Kimberly. 2011. Exploring the roles of redundancy and ambiguity in variable subject expression: A comparison of native and non-native speakers. In Claudia Borgonovo, Manuel Español-Echevarría & Philippe Prévost

- (Eds.), *Selected proceedings of the 12th Hispanic Linguistics Symposium*. Somerville, MA: Cascadilla Proceedings Project.
- Guitart, Jorge M. 1982. Conservative versus radical dialects in Spanish: implications for language instruction. *Bilingual Education for Hispanic students in the United States*, ed. by Joshua A. Fishman and Gary D. Keller, 167-77. New York: Teachers College Press
- Guy, Gregory R. 1993. The quantitative analysis of linguistic variation. *American dialect research*, ed. by Dennis R. Preston, 223–49. Amsterdam: John Benjamins.
- Guy, Gregory R. 2005. Letter to the Editor, *Language* 81.3:561-564
- Guy, Gregory R. 2007. Lexical exceptions in variable phonology. *U. Penn Working Papers in Linguistics, Volume 13.2, 2007*
- Guy, Gregory R., Jennifer Hay and Abby Walker. 2008. Phonological, lexical, and frequency factors in coronal stop deletion in Early New Zealand English. Paper presented at Laboratory Phonology 11, Wellington NZ.
- Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39: 1041-1070.
- Hay, Jennifer and Joan Bresnan. 2006. Spoken syntax: the phonetics of ‘giving a hand’ in New Zealand English. *The Linguistic Review* 23, 321-349.
- Jaeger, T. Florian, 2006. Redundancy and syntactic reduction in spontaneous speech. PhD dissertation, Stanford University.
- Kapatsinki, Vsevolod. 2009. Adversative conjunction choice in Russian (no, da, odnako): Semantic and syntactic influences on lexical selection. *Language Variation and Change* 21: 157–73.
- Lafond, Larry, Hayes, Rachel & Bhat, Rakesh. 2000. Constraint demotion and null-subjects in Spanish L2 acquisition. In Joaquim Camps & Caroline Wiltshire (Eds.), *Romance syntax, semantics and L2 acquisition: Selected papers from the 30th Linguistic Symposium on Romance Languages* (pp. 121-135). Amsterdam: John Benjamins.
- Lapidus, Naomi & Otheguy, Ricardo. 2005b. Overt nonspecific *ellos* in the Spanish of New York. *Spanish in Context*, 2, 157-176.
- Liceras, Juana M. 1989. On some properties of the “pro-drop” parameter: Looking for

- missing subjects in non-native Spanish. In Susan Gass & Jacqueline Schachter (Eds.), *Linguistic Perspectives on Second Language Acquisition* (pp.109-133). Cambridge: CUP.
- Lipski, John M. 1994. *Latin American Spanish*. London: Longman Publishers.
- Lipski, John M. 1996. Patterns of pronominal evolution in Cuban-American bilinguals. *Spanish in contact: issues in bilingualism*, ed. by Ana Roca and John B. Jensen, 159-86. Somerville, MA: Cascadilla Press.
- Montrul, Silvina & Rodriguez Louro, Celeste. (2006). Beyond the syntax of the Null Subject Parameter: A look at the discourse-pragmatic distribution of null and overt subjects by L2 learners of Spanish. In Torrens, Vincent and Linda Escobar (eds.), *The Acquisition of syntax in Romance languages*, pp. 401–418. Amsterdam: John Benjamins
- Morales, Amparo. 1997. La hipótesis funcional y la aparición del sujeto no nominal: El español de Puerto Rico. *Hispania* 80.153–65.
- Miyajima, Atsuko. 2000. Spanish subject pronoun expression and verb semantics. *Sophia Linguistica*, 46-47, 73-88.
- Myers, James and Gregory Guy. 1997. Frequency effects in variable lexical phonology', *University of Pennsylvania Working Papers in Linguistics* 4(1), 215-227.
- Morales, Amparo. 1997. La hipótesis funcional y la aparición del sujeto no nominal: el español de Puerto Rico. *Hispania* 80.153-65.
- Otheguy, Ricardo, Ana Celia Zentella & David Livert. 2007. Language and dialect contact in Spanish in New York: Towards the formation of a speech community. *Language* 83.1-33.
- Otheguy, Ricardo & Zentella, Ana Cecilia. 2007. Apuntes preliminares sobre el contacto lingüístico y dialectal en el uso pronominal del español en Nueva York. In Kim Potowski & Richard Cameron (Eds.), *Spanish in Contact: Policy, Social and Linguistic Inquiries*, pp. 275-298. Amsterdam: John Benjamins.
- Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In *Frequency and the emergence of linguistic structure*, edited by Joan Bybee. 137-158. John Benjamins.

- Pluymaekers, Mark. Ernestus, Mirjam, and Baayen, Harald R. 2005. Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*. Volume 18, Issue 4, 2561-2569.
- Phillips, Betty S. 1984. Word Frequency and the actuation of sound change. *Language* 60 (1984): 320-342.
- Phillips, Betty S. 1999. The mental lexicon: Evidence from lexical diffusion." *Brain and Language* 68: 104-109.
- Phillips, Betty S. 2001. Lexical frequency, lexical diffusion, and lexical analysis. Frequency and the emergence of linguistic structure, ed. by Joan Bybee and Paul Hopper, 4. Amsterdam: John Benjamins.
- Poplack, Shana. 1992. The inherent variability of the French subjunctive. Theoretical analyses in Romance linguistics, ed. by Christiane Laeuffer and Terrell A. Morgan, 235 – 63. Amsterdam: John Benjamins.
- Poplack, Shana. 2000. Introduction. *The English history of African American English*, ed. by Shana Poplack. Oxford: Blackwell, pp. 1-34
- Poplack, Shana. 2001. Variability, frequency, and productivity in the irrealis domain of French. Frequency and the emergence of linguistic structure, ed. by Joan Bybee and Paul Hopper, 405-428. Amsterdam: John Benjamins.
- Raymond, William D., and Esther L. Brown. 2011. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. *Topics in Linguistics Studies and Monographs*, ed. by Stefan Th. Gries & Dagmar S. Divjak. Mouton de Gruyter, to appear.
- Serrano, María José. 1996. El análisis del discurso en variación sintáctica. *Hispanic Linguistics*, 8, 154-177.
- Schwenter, Scott & Torres Cacoullos, Rena. 2008. Defaults and interdeterminacy in temporal grammaticalization: The ‘perfect’ road to perfective. *Language Variation and Change*, 20, 1-39.
- Silva-Corvalán, Carmen. 1982. Subject expression and placement in Mexican-American Spanish. In Jon Amastae & Lucía Elías-Olivares (Eds.), *Spanish in the United States: Sociolinguistic aspects* (pp. 93-120). New York: Cambridge University Press.

- Silva-Corvalán, Carmen. 1994. *Language contact and change: Spanish in Los Angeles*. Oxford: Oxford University Press.
- Silva-Corvalán, Carmen. 1995. The study of language contact: An overview of the issues. *Spanish in four continents: Studies in language contact and bilingualism*, ed. by Carmen Silva-Corvalán, 3–14. Washington, DC: Georgetown University Press.
- Silva-Corvalán, Carmen. 1997a. Avances en el estudio de la variación sintáctica: La expresión del sujeto. *Cuadernos del Sur* 27.35–49.
- Silva-Corvalán, Carmen. 1997b. Variación sintáctica en el discurso oral: Problemas metodológicos. *Trabajos de sociolingüística hispánica*, ed. by Francisco Moreno Fernández, 115–35. Alcalá de Henares: Universidad de Alcalá.
- Silva-Corvalán, Carmen. 2001. *Sociolingüística y pragmática del español*. Washington, D.C.: Georgetown University Press.
- Torres Cacoullos, Rena, and Fernanda Ferreira. 2000. Lexical frequency and voiced labiodental-bilabial variation in New Mexican Spanish. *Southwest Journal of Linguistics* 19.2:1-17.
- Torres Cacoullos, Rena and Catherine E. Travis. 2010. Variable *yo* expression in New Mexico: English influence? Spanish of the U.S. southwest: A language in transition, ed. by Susana Rivera-Mills and Daniel J. Villa, 185-206. Madrid: Iberoamericana.
- Torres Cacoullos, Rena, and James A. Walker. 2009. On the persistence of grammar in discourse formulas: a variationist study of that. *Linguistics* 47.1:1-43.
- Tottie, Gunnel. 1991. Lexical diffusion in syntactic change: frequency as a determinant of linguistic conservatism in the development of negation in English. *Historical English syntax*, ed. by D. Kastovsky, 439-67. Berlin: Mouton de Gruyter.
- Travis, Catherine. The *yo-yo* effect: Priming in subject expression in Colombian Spanish. *Selected papers from the 34th Linguistic Symposium on Romance Languages (LSRL), Salt Lake City, 2004*, ed. by Randall Gess and Edward J. Rubin, 329–49. Amsterdam: John Benjamins.
- Travis, Catherine E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19.101–35.

- Travis, Catherine E. and Agripino S. Silveira. 2009. The role of frequency in first-person plural variation in Brazilian Portuguese: Nós vs. a gente. *Studies in Hispanic and Lusophone Linguistics* 2.2.347-376.
- Verhagen, Arie. 2006. Syntax, recursion, productivity - a usage-based perspective on the evolution of grammar. A Lubotsky, J. Schaeken, J. Wiedenhof (eds.), *Evidence and Counter-Evidence*, Festschrift F. Kortlandt, Volume I (SSGL 32 and 33). Amsterdam - New York.
- Vitevitch, M. S., & Luce, P. A. 1998. When words compete: Levels of processing in perception of spoken words. *Psychological Science* 9.325–329.
- Vitevitch, M.S. 2002. Naturalistic and experimental analyses of word frequency and neighborhood effects in slips of the ear. *Language and Speech*, 45, 407-434.
- Vitevitch, M. S., & Rodriguez, E. 2005. Neighbourhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders* 3.64–73.
- Watson, Kevin, and Lynn Clark, 2010. t-to-r in British English: frequency, constructions, and sociolinguistics. paper presented at NWAV-39, University of Texas at San Antonio.

## Notes

---

<sup>1</sup> While we discuss the effects of several of these factors in the methodology section, it is not our focus to provide a detailed examination of all of them. For further depth on this topic we refer the reader to Otheguy et al. 2007.

<sup>2</sup> Although this study is the first to examine *morphological regularity* as a constraint on pronoun use, it should be noted that analyzing SPP use in light of the morphological regularity of verb forms is not entirely novel in spirit. Indeed, most studies that investigate TMA and SPP use suggest, implicitly or explicitly, that the relatively more regular TMA forms like the imperfect indicative are associated with more SPP use.

<sup>3</sup> Regular vocalic alternations in which stressed vowels are diphthongized or raised, as in *poder* → *puede* ‘she can’, *tener* → *tiene* ‘he has’, and *decir* → *dice* ‘she says’ were not treated as irregulars in the present analysis. This decision is based on the relatively large fraction of forms that are derived by these processes. They are systematic, highly frequent and robustly productive throughout the Spanish verbal inflectional system and beyond. To the extent that they are thus ‘regular’, it is preferable to treat them as such.

<sup>4</sup> As noted, we counted surface forms, which for certain forms are equivalent across certain person/number combinations (e.g. *era* is both first and third person singular), or between lexical roots (*fui* is the first singular preterite form of both *ser* ‘to be’ and *ir* ‘to go’). However, in our quantitative analysis, person/number is treated as a conditioning factor on pronoun use.

<sup>5</sup> The results presented here focus on the linguistic constraints on this process. As noted above, the speaker sample we studied included two nationality groups – Mexicans and Dominicans – and two levels of residence in the US – people born and raised in New York vs. those newly arrived as immigrants. We show in section 3.4 that the results presented here are consistent across these social subdivisions.

<sup>6</sup>A similar result is obtained if words are treated in terms of their rank order of frequency:  $r(4,916) = -.035$ ,  $p < .02$ . The negative correlation statistic for rank reflects the fact that rank orders run the opposite direction from frequencies – the highest frequency item occurs 208 times but has a rank of 1.

<sup>7</sup> This analysis excludes four TMA combinations with low Ns in our corpus: the future indicative (N=15), the conditional (N=29), and the past (N=65) and perfect subjunctives (N=3). The full figures for all TMA combinations are given in Table 5a here.

	N Verbs	% Overt Pronouns
Present Indicative	2,695	36
Preterite Indicative	877	29
Imperfect Indicative	708	43
Periphrastic Future	140	36
Future Indicative	15	40
Conditional	29	17
Present Subjunctive	96	31
Past Subjunctive	65	34
Imperative	110	1
Perfect Indicative	176	33
Perfect Subjunctive	3	33
F(10, 4904) = 8.993 , $p < .001$		

<sup>8</sup> These effects persist across different persons and numbers.

<sup>9</sup> The analysis of person/number presented here is restricted to singular forms, because all plural forms fell in our infrequent category, making a frequent vs. infrequent comparison impossible.

<sup>10</sup> As was the case with person/number, not all possible TMA values are considered because many of them are only represented by infrequent forms, thus making a frequent vs. infrequent comparison impossible.

---

<sup>11</sup> Similar results are obtained when raw frequency is substituted for discrete frequency: this frequency measure also fails to be a significant predictor of SPP use, with a Wald value of .37,  $p = .543$ .