

2023

Contextual frequency and morphosyntactic variation: an exemplar-theoretic variationist analysis of Spanish subject pronouns

<https://hdl.handle.net/2144/49365>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**CONTEXTUAL FREQUENCY AND
MORPHOSYNTACTIC VARIATION: AN
EXEMPLAR-THEORETIC VARIATIONIST ANALYSIS
OF SPANISH SUBJECT PRONOUNS**

by

DANIELLE DIONNE

A.A., Palm Beach State College, 2013
B.A., Florida Atlantic University, 2015
M.A., Florida Atlantic University, 2017
M.A., Boston University, 2021

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2023

© 2023 by
DANIELLE DIONNE
All rights reserved

Approved by

First Reader

Daniel Erker, PhD
Associate Professor of Spanish and Linguistics

Second Reader

Neil Myler, PhD
Associate Professor of Linguistics

Third Reader

Rafael Orozco, PhD
Professor of Linguistics and Spanish
Louisiana State University

Acknowledgments

People from all corners of my life have contributed to my completion of this dissertation. First and foremost, I extend immense gratitude to Daniel Erker. Dr. Erker has been my advisor since I first began my doctoral studies in Fall of 2018. I have learned so much from him in the classroom as a student, in the lab as a researcher, but also in academia as his advisee. In all of these contexts, Dr. Erker was ready with a car metaphor or Big Lebowski quote for encouragement. Without his supportive guidance and constructive feedback, I would not have been able to complete this project to the level that it is today.

I would also like to thank Neil Myler and Rafael Orozco, who generously provided their time and wisdom over the course of this project. Their excitement and expertise made this project feel achievable even in the thick of it. I'd like to thank Ricardo Otheguy and Ana Celia Zentella, who created the New York Corpus and have generously shared it with students like me. My gratitude extends to all of the informants in the New York Corpus and the Boston Corpus, and to the research assistants and students that helped with all aspects of data collection.

Several other individuals have supported me and provided encouragement throughout this process. I am grateful to my colleagues and friends in the Department of Linguistics, who have trudged through with me as we moved through graduate studies during a global pandemic. Lee-Ann Vidal Covas, Alex Kohut, and Brady Dailey played a particularly important role as my graduate school community over the last five years. Their friendship and collaboration have been crucial to getting me through.

Finally, I want to thank my family and friends outside of academia. I am deeply grateful to my parents, Terrie and Raynald, and my sister, Juliane, who have cheered me on in all of my endeavors. Their unconditional love and support have helped me reach this great achievement. I am thankful to Deanna, Chantal, and Darnella for

being the best hype-women I could ask for. They have rallied behind me through it all and have never missed an opportunity to laugh, cry, or dance with me. To those family, friends, and linguists that I have not explicitly addressed here, please know that I am thankful to you as well! In the wise words of a previous professor, “it takes a village to earn a Ph.D.”, and I am immensely grateful for my village.

CONTEXTUAL FREQUENCY AND
MORPHOSYNTACTIC VARIATION: AN
EXEMPLAR-THEORETIC VARIATIONIST ANALYSIS
OF SPANISH SUBJECT PRONOUNS

DANIELLE DIONNE

Boston University, Graduate School of Arts and Sciences, 2023

Major Professor: Daniel Erker, PhD

Associate Professor of Spanish and Linguistics

ABSTRACT

This study incorporates insights from Usage Based Grammar (UBG) into variationist research on morphosyntactic variation in Spanish. Specifically, this dissertation investigates the impact on pronoun use of lexical frequency, or the number of times a finite verb appears in a large data set based on spontaneous speech from 221 speakers in two locales (New York City, NY and Boston, MA), as well as a series of context-based frequency metrics in a Variationist study of Spanish Subject Personal Pronoun (SPP) variation (e.g. *Yo creo* vs. *creo* ‘I think’). This investigation elucidates the nature of frequency effects (both lexical and contextual) on pronoun use *and* on the other linguistic factors that have been shown to impact pronoun use. Through this investigation, this dissertation is able to draw conclusions on the nature of linguistic variation and make inferences surrounding the mental representations underlying sociolinguistic patterns.

In the past, frequency has been investigated in subject pronoun production as it pertains to the rate of the finite verb, with researchers counting the instances of each

verb's occurrence within a corpus. This approach has produced mixed results. One study has shown that frequency modulates or amplifies the effects of other linguistic predictors, providing evidence that suggests lexical frequency does not directly impact pronoun use in a uniform or monotonic way (Erker and Guy, 2012). A few studies have replicated some version of these modulating effects, though they have not found as consistent amplification effects across linguistic constraints. Other studies have found contradictory frequency effects, showing only a main effect of frequency (high frequency corresponding to high pronoun use in some studies and low pronoun use in others) with no amplification effects, or no frequency effects at all. Further, Usage Based Grammar frameworks, which are often referenced in studies exploring lexical frequency, posit that speakers are not only sensitive to the rate of use of linguistic forms, but also the detailed contexts in which these forms appear. Such "rich memories", as they are referred to in UBG, are said to constitute the mental representations of these forms.

The mixed results in the literature, together with the UBG notion of rich memories, motivate the current study, which investigates the relationship between contextual frequency and pronoun use, since contextual frequency metrics (as opposed to overall frequency) might shed more light on frequency effects in morphosyntactic variation. The contextual frequency metrics analyzed in the current dissertation consist of the frequencies at which finite verbs appear in four combinations of the factor values of two variables, referred to as Switch Reference (i.e. whether the previous verb has a different referent or the same referent as the target site of variation) and Preceding Pronoun (i.e. whether the immediately preceding site of pronominal variation has a pronoun present or absent). The four combinations on which contextual frequency metrics are based are therefore: (1) 'Different Referent/Preceding Pronoun Present', (2) 'Different Referent/Preceding Pronoun Absent', (3) 'Same Referent/Preceding

Pronoun Present’, or (4) ‘Same Referent/Preceding Pronoun Absent’.

Analysis of 88,001 tokens of pronominal presence or absence generally replicate the modulating effects of overall verb frequency observed by Erker and Guy (2012), i.e. the effects of several linguistic factors are amplified for frequent verb forms. Moreover, the analysis of contextual frequency reveals that verb forms must reach a certain overall frequency threshold in order for contextual properties to impact pronoun use. This finding aligns with the UBG prediction that the most frequent context in which a verb appears will dominate the overall pronominal tendencies of the verb, as long as that verb is sufficiently frequent in discourse. Overall, this study concludes that the linguistic variation observed in language use aligns with the usage-based approach that contextual frequency effects accumulate in the mental representations that underlie sociolinguistic patterns.

Contents

1	Introduction	1
1.1	Variationist Sociolinguistics and Usage-Based Grammar	2
1.2	Challenges to Incorporating Usage-Based insights into Variationist Research	7
1.3	Spanish Subject Pronoun Use and Usage-Based Grammar	12
1.4	Research Questions	16
1.5	Broad Predictions and Overview of the Results	16
1.6	Dissertation Structure	18
2	Review of Existing Literature	20
2.1	Variationist Sociolinguistics	20
2.2	Variationist Sociolinguistic Research on Spanish Subject Pronouns	25
2.2.1	Linguistic Constraints on Variable Pronoun Production	26
2.2.2	Social factors and Their Impact on Spanish Subject Pronouns	31
2.2.3	The Role of Frequency in Spanish Subject Pronouns	34
2.3	Usage-Based Grammar	38
2.4	Better Representing Usage-Based Insights in Variationist Research	44
2.4.1	Contextual Frequency Metrics Proposed for the Present Study	45
3	Methodology & Predictions	49
3.1	Overview	49
3.2	The Corpus	50
3.2.1	Sociolinguistic Interviews	50

3.2.2	The Speakers	50
3.3	The Dependent Variable	53
3.4	The Independent Variables	54
3.4.1	Linguistic Variables	54
3.4.2	Lexical Frequencies	57
3.4.3	Contextual Frequencies	58
3.4.4	Summary of the Independent Variables	62
3.5	Predictions for Contextual Frequencies	63
3.6	Statistical Methods	66
3.6.1	Replicating Erker and Guy (2012)	66
3.6.2	Extending Erker and Guy (2012)	67
4	Replicating Erker & Guy (2012): Return to Lexical Frequency	69
4.1	General Overview of the Data	70
4.2	Main Effects of Lexical Frequency	74
4.3	Morphological Regularity	82
4.3.1	Interaction with Discrete Frequency	83
4.4	Person and Number	86
4.4.1	Interaction with Discrete Frequency	87
4.5	TMA	91
4.5.1	Interaction with Discrete Frequency	94
4.6	Semantic Category	99
4.6.1	Interaction with Discrete Frequency	100
4.7	Switch Reference	102
4.7.1	Interaction with Discrete Frequency	103
4.8	Summary of Replication Results	105
4.9	Exploring Additional Linguistic Constraints	110

4.9.1	Preceding Pronoun	110
4.9.2	Morphological Regularity 2.0	114
4.9.3	Semantic Category 2.0	117
4.10	Summary of All Linguistic Predictors	122
5	Expanding Erker & Guy (2012): Analysis of Contextual Frequency Metrics	123
5.1	General Description of Contextual Frequency	124
5.2	Relationship Between Lexical and Contextual Frequency	127
5.3	Log Contextual Frequencies	129
5.4	Exploring Contextual Frequency Non-Linearly	131
5.5	Favoring Context Ratio	134
5.6	Disfavoring Context Ratio	142
6	Discussion & Conclusions	148
6.1	The Nature of the Replication Results	148
6.2	The Incorporation of Contextual Frequency Metrics	152
6.3	The Theoretical Implications of Incorporating UBG Insights into Variationist Research	153
6.4	Testing for Pronominal Prefabrication	155
6.5	Further Considerations	161
6.5.1	Different methodologies for frequency	161
6.5.2	Investigating social factors	163
6.6	Other Avenues for Future Analyses	164
6.7	Conclusions	165
A	Appendix	167
A.1	Speaker Demographics	167
A.1.1	Regional Origin	167

A.1.2 Arrival to the U.S.	170
References	172
Curriculum Vitae	183

List of Tables

3.1	Regional origin for Speakers in the Boston Corpus, NYC Corpus, and the joint OZC-BSC Corpus.	52
3.2	Examples illustrating the novel, context-dependent variable: CONTEXTUAL FREQUENCY.	59
3.3	Summary of all Independent Variables	64
4.1	Frequent verb forms	73
4.2	Mean raw frequency: frequent vs. infrequent forms	74
4.3	Morphological regularity of the verb	82
4.4	Main effect of Morphological Regularity; Model configuration: <code>glmer(Pronoun ~ Morphological Regularity + (1 Verb))</code> ; Reference level: Irregular	83
4.5	Interaction between Morphological Regularity and Discrete Frequency; Model configuration: <code>glmer(Pronoun ~ Regularity*Discrete Frequency)</code> ; Reference levels: Irregular, Frequent, Irregular*Frequent	85
4.6	Person and number of the verb, all combinations considered	86
4.7	Main effect of Person/Number; Model configuration: <code>glmer(Pronoun ~ Person/Number + (1 Verb))</code> ; Reference level: First Plural	87
4.8	Person/Number & Discrete Frequency Interaction; Model configuration: <code>glmer(Pronoun ~ Person/Number*Discrete Frequency + (1 Verb))</code> ; Reference levels: First plural, Frequent, First plural*Frequent	90
4.9	TMA of the verb, all combinations considered	92

4.10	Main effect of TMA; Model configuration: <code>glmer(Pronoun ~TMA + (1 Verb))</code> ; Reference level: Conditional	93
4.11	Main effect of top three factor values of TMA; Model configuration: <code>glmer(Pronoun ~ TMA + (1 Verb))</code> ; Reference level: Imperfect indicative	94
4.12	Interaction between a subset of TMA & Discrete Frequency; Model configuration: <code>glm(Pronoun ~TMA*Discrete Frequency)</code> ; Reference levels: Imperfect indicative, Frequent, Imperfect indicative*Frequent	96
4.13	Interaction between TMA & Discrete Frequency; Model configuration: <code>glm(Pronoun ~TMA*Discrete Frequency)</code> ; Reference levels: Conditional, Frequent, Conditional*Frequent	97
4.14	Semantic content of the verb	99
4.15	Main effect of semantic category; Model configuration: <code>glmer(Pronoun ~Semantic category + (1 Verb))</code> ; Reference level: External activity verbs	100
4.16	Interaction between Semantic Category and Discrete Frequency; Model configuration: <code>glmer(Pronoun ~Semantic category*Discrete Frequency + (1 Verb))</code> ; Reference levels: External activity verb, Frequent, External activity verb*Frequent	102
4.17	Switch Reference	103
4.18	Main effect of Switch Reference; Model configuration: <code>glmer(Pronoun ~ Switch Reference + (1 Verb))</code> ; Reference level: Different referent	103
4.19	Interaction between Switch Reference and Discrete Frequency; Model configuration: <code>glmer(Pronoun ~ Switch Reference*Discrete Frequency + (1 Verb))</code> ; Reference levels: different referent, frequent, different referent*frequent	105

4.20	Summary of main effects, interactions with frequency, and evidence of amplification effects for core constraints	106
4.21	Summary of marginal and conditional R^2 for the main effect and interaction models of core constraints	107
4.22	Preceding Pronoun	111
4.23	Main effect of Preceding Pronoun; Model configuration: <code>glmer(Pronoun ~ Preceding Pronoun + (1 Verb))</code> ; Reference level: preceding pronoun absent	111
4.24	Interaction between Preceding Pronoun and Discrete Frequency; Model configuration: <code>glmer(Pronoun ~ Preceding Pronoun*Discrete Frequency + (1 Verb))</code> ; Reference levels: preceding pronoun absent, frequent, preceding pronoun absent*frequent	113
4.25	Morphological Regularity 2.0	114
4.26	Main effect of Morphological Regularity 2.0; Model configuration: <code>glmer(Pronoun ~ Morphological Regularity 2.0 + (1 Verb))</code> ; Reference level: irregular	115
4.27	Interaction between Morphological Regularity 2.0 and Discrete Frequency; Model configuration: <code>glmer(Pronoun ~ Morphological Regularity 2.0*Discrete Frequency + (1 Verb))</code> ; Reference levels: irregular, frequent, irregular*frequent	117
4.28	Semantic Category 2.0	118
4.29	Main effect of Semantic Category 2.0; Model configuration: <code>glmer(Pronoun ~ Semantic Category 2.0 + (1 Verb))</code> ; Reference level: estimative	118

4.30	Interaction between Semantic Category 2.0 and Discrete Frequency; Model configuration: <code>glmer(Pronoun ~ Semantic Category 2.0*Dis- crete Frequency + (1 Verb))</code> ; Reference levels: estimative, frequent, estimative*frequent	121
4.31	Summary of main effects, interactions with frequency, and evidence of amplification effects for all linguistic constraints	122
5.1	Contextual Frequency metrics	124
5.2	Correlation matrix for Log Frequency and four Contextual Frequency Metrics	128
5.3	Correlation matrix of pronoun rates for the four Contextual Frequency Metrics	131
5.4	Summary of correlations between FCR and pronoun rate for low log frequency and high log frequency groups and correlations between log frequency and pronoun rate for low FCR and high FCR groups. . . .	138
5.5	Descriptive information for forms in the high log frequency group or- ganized by FCR	140
5.6	Interaction between FCR and Log Frequency; Model configuration: <code>glmer(Pronoun ~ FCR*Log Frequency + (1 Verb))</code> ; Reference level: 0.00 FCR, 0.00 Log Frequency, 0.00 FCR*0.00 Log Frequency	141
5.7	Summary of correlations between Log frequency and pronoun rate for low DCR and high DCR groups and correlations between DCR and pronoun rate for low log frequency and high log frequency groups. . .	143
5.8	Interaction between DCR and Log Frequency; Model configuration: <code>glmer(Pronoun ~ DCR*Log Frequency + (1 Verb))</code> ; Reference level: 0.00 DCR, 0.00 Log Frequency, 0.00 DCR*0.00 Log Frequency	146

6.1	Summary of high frequency forms and their randomly sampled equivalents; Model configurations: <code>glmer(Pronoun ~ Switch Reference + Preceding Pronoun + (1 Speaker))</code>	158
A.1	Age and Sex for Speakers in the Boston Corpus, NYC Corpus, and OZC-BSC joint corpus.	168
A.2	Regional origin and Country of origin for Speakers in the Boston Corpus, NYC Corpus, and Overall.	169
A.3	Age of arrival (age) and PLUS for Speakers in the Boston Corpus, NYC Corpus, and OZC-BSC Corpus.	171

List of Figures

4·1	Zipf’s law illustrated through raw frequency distribution of verb forms in the OZC-BSC Corpus.	71
4·2	Percent pronouns present within each raw frequency	75
4·3	Percent pronouns present within each log frequency	77
4·4	Frequent and Infrequent forms by percent SPPs present	78
4·5	Percent SPPs present for highest-frequency forms	79
4·6	Percent SPPs present for verb forms that appear at least 440 times in the corpus (comprising 0.5% of the corpus)	81
4·7	Morphological regularity: frequent vs. infrequent forms	84
4·8	Subset of Person/Number: frequent vs. infrequent forms	88
4·9	Person/Number: frequent vs. infrequent forms, all forms considered	89
4·10	Subset of TMA: frequent vs. infrequent forms	95
4·11	TMA: frequent vs. infrequent forms, all forms considered	98
4·12	Semantic content: frequent vs. infrequent forms	101
4·13	Switch reference: frequent vs. infrequent forms	104
4·14	Amplification effects of Discrete Frequency	108
4·15	Preceding Pronoun: frequent vs. infrequent forms	112
4·16	Morphological Regularity 2.0: Frequent vs. infrequent forms	116
4·17	Semantic Category 2.0: Infrequent and frequent forms	120
5·1	Log Contextual Frequencies and Percent Pronouns Present	130

5.2	Percent pronouns present by Log Frequency of appearing in a pronoun disfavoring context for three different lexical frequency groups	132
5.3	Favoring Context Ratio and percent pronouns present	135
5.4	Favoring Context Ratio (rounded to two decimal points) and percent pronouns present	136
5.5	Pronoun rate by FCR for log-controlled groups	139
5.6	Disfavoring Context Ratio and percent pronouns present	142
5.7	Pronoun rate by DCR for log-controlled groups	145
6.1	Percent pronouns present within each log frequency (accounting for token and type frequency)	150

List of Abbreviations

AIC	Akaike Information Criterion
CSD	Coronal Stop Deletion
DCR	Disfavoring Context Ratio
FCR	Favoring Context Ratio
OZC-BSC	Otheguy-Zentella & Boston Spanish Joint Corpus
R_c^2	Conditional R^2
R_m^2	Marginal R^2
SPP	Subject Personal Pronoun
UBG	Usage-Based Grammar

Chapter 1

Introduction

The so-called *Chomskyan Revolution*, one of the major (perhaps *the* major) inflection points in modern linguistic inquiry, began in the late 1950s and early 1960s (Chomsky, 1957, 1965). As a direct response to Behaviorism (Skinner, 1957), Chomsky proposed a theory of human language that is rooted in generativity – that humans can produce and understand novel utterances without ever previously encountering them. He notes this as a unique component of human language, as opposed to other animal communication systems, that cannot be explained via direct cause-and-effect relations. Instead, Chomsky argues that humans are born with an innate language mechanism –called *Universal Grammar*– that includes built-in (syntactic) operations and allows for the acquisition of linguistic structure. He set forth a framework for linguistic inquiry that is most concerned with the ways in which all human languages are homogeneous, outlining abstract grammars, or mental representations of language in the mind, that are modular in nature (Chomsky and Halle, 1968). Further, Chomsky emphasizes that his approach to linguistic inquiry prioritizes linguistic competence (i.e. knowledge about a language) and purposefully excludes linguistic performance (i.e. actual production of that language). He affirms, “the generative grammar that expresses the speaker-hearer’s knowledge of the language... does not, in itself, prescribe the character or functioning of a perceptual model or a model of speech production” (Chomsky, 1965, p.9). In other words, the generativist endeavor is not interested in language use. By excluding all aspect of language use (percep-

tion and production) from the structural rules that apply to them, the generativist grammar is invariant in application and algorithmic in nature.

Around the same time as the rise of Generative Linguistics and shortly thereafter, however, other scholars, convinced of the necessity to study rather than abstract away from language use, were laying the foundation for other paradigms to linguistic inquiry. Two such schools of thought are known today as *Variationist Sociolinguistics* and *Usage-Based Theories of Language*. It is with these two fields that the present dissertation enters into conversation, with particular focus on their shared emphasis on patterns observed in actual linguistic behavior.

1.1 Variationist Sociolinguistics and Usage-Based Grammar

What is now known as Variationist Sociolinguistics came about in the 1960s, and the field has evolved over the course of decades in how variation in language use is analyzed and what conclusions are drawn from these analyses (Labov, 1963, 1966; Weinreich et al., 1968; Wolfram, 1991; Labov, 2001, 2011). The main objectives in Variationist Sociolinguistics are to understand the systematic variability that is present in human language, to determine how this variability interacts with and ultimately drives language change, and to situate this variability and change within a social context. Variationists provide quantitative descriptions of language use, which are hypothesized to reflect underlying linguistic knowledge. These quantitative descriptions of language (i.e. quantitative models of language use proposed to reflect the mental grammar) are positioned within social and linguistic contexts. That is to say that the linguistic performance of a group of individuals is described relative to the community that the individuals belong to.

Variationists attempt to understand usage through the lens of the sociolinguistic variable, which is defined as “two or more ways of saying the same thing” (Labov,

1972b, p. 188). Let's consider the sociolinguistic variable of coda /ɹ/ (e.g. /ɹ/ in *store*). In some dialects of English such as Boston English, coda /ɹ/ has two variants: (1) r-ful, and (2) r-less. When variationists investigate how and when these variants are produced, they uncover sets of competing constraints on the realization of coda /ɹ/. From the very beginning of sociolinguistic research (Labov, 1966), these constraints have included linguistic (e.g. phonetic context, morpheme status, or speech rate) and social factors (e.g. age, sex, socioeconomic status). In the context of coda /ɹ/, variationist research has shown that the deleted variant appears more often before a consonant, in free morphemes, and in the middle of a word (Stanford, 2019). In contrast, the realized variant appears more often before a vowel or a pause, in bound morphemes, and at the end of a word. As for social constraints, research suggests that older Bostonians are less likely to produce /ɹ/, while younger Bostonians favor the /ɹ/ (Stanford, 2019). The effect that these competing constraints have on variable usage is extremely organized and directly observable. Variationist research clearly demonstrates that the distribution of variants is dependent upon linguistic and social factors (i.e. constraints), and much of the most recent research on sociolinguistic variation aims to uncover additional constraints on sociolinguistic variables¹.

Another school of linguistic thought devoted to studying language use is referred to today as Usage-Based Grammar (UBG), or Exemplar Theory². UBG emerged in the 1980s in response to Generativist theories surrounding mental grammar (Bybee, 1985, 1998; Pierrehumbert, 2001; Bybee, 2001, 2006, 2010). Addressing the generative the-

¹One open line of inquiry within the field of Sociolinguistics is related to the issue of categorizing constraints in the traditionally binary fashion of “Social” and “Linguistic” constraints. Some linguists have categorized constraints as “Internal” and “External”, which potentially mediates the challenges that arise for constraints that are not purely linguistic in nature (Labov, 1994). The categorization of constraints will be discussed in further detail in Chapter 6.

²In the current dissertation, the terms Exemplar Theory and UBG will be used interchangeably. Though these two research traditions are not identical, Exemplar Theory is considered to fall under the umbrella of UBG-based theories (Pierrehumbert, 2001; Bybee, 2010). Exemplar Theory is a method for representing a grammar whose structure emerges through use. In other words, Exemplar Theory offers a way to model a usage-based grammar.

ory on mental grammar, Bybee states “ grammar is not static and rigid, as predicted by the innatist position. Rather, grammar is constantly changing: old grammatical constructions are constantly being replaced with newly-formed constructions” (Bybee, 1998, p. 250). Proponents of UBG, unlike those of the Chomskyan approach, argue *against* the separation of linguistic structure and language use. Instead, they envision grammar as a complex web containing detailed information from speakers’ linguistic experiences that evolves over the course of a speaker’s lifetime. Linguistic experiences create rich memories (i.e. remembered experiences that contain detailed linguistic and social information surrounding the experience) of specific forms and the collections of such memories constitute all levels of linguistic knowledge (e.g. phonological, morphological, syntactic, etc.). Grammatical structure is viewed as emergent in nature; it arises through language use, and is categorized and organized by speakers through general cognitive properties (pattern recognition, for example) as opposed to innate language specific capacities. Unlike the Generative approach, which asserts that sentential structure exists separate from lexical meaning, UBG argues that there is no separation of structural form from semantic information. Instead, structure and meaning are closely connected within the mental grammar. With each linguistic experience, new rich memories – or exemplars – are added to “exemplar clouds” (i.e. groupings of exemplars of similar form), thereby increasing the complexity and robustness of a speaker’s mental representation of that form. The exemplar cloud itself contains a spectrum of variation for that form that the speaker has experienced over the lifespan.

UBG asserts that language users are not only remembering/storing instances of language use across the lifespan, but they are also using pattern recognition and other cognitive properties to classify/categorize them. Exemplar Theory provides a model for how this process is generally representable in the mind. Pierrehumbert (2001), in

a seminal contribution to UBG theorizing, explains that each experience of a form is categorized into its relevant exemplar cloud based on a series of features. Features are thought to include linguistic and social information such as formant estimates, verb tense, or speaker. The form is essentially graded based on the features it has and how closely it aligns to the features of existing exemplar clouds. The exemplar cloud whose features most-closely resemble the features of the experienced form is then chosen by the language user for classification. If the form does not resemble exemplars in existing exemplar clouds, it is stored on its own. Once the exemplar cloud that most closely matches the form is selected, the language user then compares the token to existing exemplars within that exemplar cloud that vary along multiple lines. If a new exemplar possesses features that are near-identical to an existing exemplar or group of exemplars, then it is added to that group within the exemplar cloud. As exemplars are added to a given exemplar cloud, the details of use for that exemplar are said to increase. Thus, in UBG, higher frequency forms have more detailed mental representations, since each form is stored.

Frequency of use also plays a role in the resting activation level of exemplars within a given exemplar cloud. Exemplars that are more frequent and/or were encountered more recently are said to have higher resting activation levels than exemplars that are less frequent or are tied to older memories (Pierrehumbert, 2001). This resting activation level contributes to the organization of exemplars in a given exemplar cloud. More similar forms are clustered closer together in the mental representation, and exemplars with higher resting activation have an advantage over less frequent/less recent exemplar in the same area. Within exemplar theoretic models of language, frequency is not explicitly coded for in the mental representation. Instead, frequency is built into the system. This conception of grammars effectively guarantees that the frequency properties of linguistic forms will matter, since exemplars with different fre-

quencies are represented differently in the mind (i.e. more or less robustly and with increasing/decreasing levels of activation). What’s more, the effects of these frequency differences have been shown to materialize in different ways across different levels of linguistic structure: highly frequent forms lean toward phonetic reduction (Bybee, 2001; Pierrehumbert, 2001; Brown, 2015), but often resist changes at the morphosyntactic level (Bybee, 1985, 2010). Extremely frequent forms can even evolve into what UBG refers to as “prefabs”, which are highly frequent multi-word expressions that become conventionalized (e.g. *X drives me Y* in Bybee, 2010).

Evolving out of similar circumstances, these two paradigms of linguistic inquiry, that is, the Variationist and Usage-Based-Grammar/Exemplar-Theoretic approaches, have some similarities: one of their main sources of dissatisfaction with the Chomskyan paradigm lies in his description of the ideal speaker-hearer possessing a homogeneous grammar comprised purely of invariant abstraction. Further, variationists and usage-based grammarians have argued that the generative method of using acceptability judgements is insufficient for fully illuminating the nature of human language. From a variationist perspective, “intuition is less regular and more difficult to interpret than speech” (Labov, 1972a, p.199). Variationist Sociolinguistics and UBG take issue with the generative approach because it argues that language use is too complex to account for and excludes usage *by design*. Given this shared discomfort and commitment to investigating usage, these two paradigms have unsurprisingly come into contact in a number of ways. One of the most direct ways, and the most relevant for the current thesis, in which these two approaches to linguistic inquiry have come into contact is on the topic of frequency.

Research in UBG has consistently shown support for the notion that frequency affects usage. In light of this, recent variationist research has explored multiple other constraints that condition the underlying structure of variant choice (Tamminga et al.,

2016). Many of these constraints consist of frequency metrics defined in various ways, and much of this research aims to understand how frequency affects sociolinguistic variation and how UBG insights can translate into variationist research. However, this research history has many instances of confounding or confusing findings, leaving multiple questions still unresolved surrounding the nature of frequency effects and UBG insights more broadly in variationist research. These questions are discussed briefly in the following section.

1.2 Challenges to Incorporating Usage-Based insights into Variationist Research

The research history described above leaves the following questions still unresolved: (1) How can insights from UBG be best leveraged in the context of the study of linguistic variation, especially outside of the domain of phonetics/phonology?; (2) What is the best way to define and describe frequency as it pertains to Variationist research?; (3) How does the UBG concept of a prefab align (or not) with sociolinguistic research?; (4) Do frequency-based constraints constitute a different kind of potential constraint on variation than constraints like Age or morphemic status? Each of these questions are considered to varying degrees in the current thesis, and they are described in more detail below.

Perhaps the largest unanswered questions surrounding frequency effects and UBG in Variationist research is how we can understand the role of frequency outside of the domain of phonology. The link between frequency and variation, though not entirely obvious or without debate, is much more straightforward in the phonological domain than in variation at the level of morphosyntax. Increased frequency contributes to an increasingly robust mental representation, which improves lexical access and perception, but also impacts muscle memory and articulatory effort. Higher frequency

forms are produced more often *and* heard more often, which has been shown to promote a reduction of the articulatory effort needed to produce a form in a manner that is understood by the listener (Bybee, 2010). These cumulative frequency effects can result in phonetic reduction, which often presents itself variably and can lead to phonological change. For this reason, the majority of investigations of frequency and attempts to incorporate usage-based insights into Variationist models of language have focused on sound phenomena (Myers and Guy, 1997; Berkenfield, 2001; Brown and Torres Cacoullos, 2003; Labov, 2006; Abramowicz, 2007; Díaz-Campos and Gradoville, 2011; Raymond and Brown, 2012; Brown and Raymond, 2012; Bongiovanni, 2014; Tamminga, 2014; Brown, 2015; Hay et al., 2015; Escalante, 2016; Todd et al., 2019; Brown et al., 2021; Purse et al., 2022, and others).

One of the earliest studies to do so is Myers and Guy (1997). In their study of coronal stop deletion, they investigate the extent to which lexical frequency impacts the rate of deletion for monomorphemic (e.g. *lift*) and past tense forms (e.g. *laughed*). They find an effect of frequency for the rate of coronal stop deletion in monomorphemic forms such that high frequency monomorphemic forms favor deletion, but they do not find any frequency effects on deletion in past tense forms. Though investigating frequency in the context of phonological variation was the most natural place to start, challenges arise outside of phonology, where the notion of frequency isn't obviously connectable to the physical processes of speech production (which are clearly frequency-biased).

Further, it remains unclear how to define and describe frequency in the first place. Many studies that have investigated frequency effects on variation have done so using different definitions of frequency. Myers and Guy (1997), for example, utilizes frequency information taken from a written corpus. Other studies have measured frequency within their corpus of study (Erker and Guy, 2012; Linford and Shin, 2013).

Some scholars have operationalized frequency as a binary categorical variable (e.g. “frequent” or “infrequent”) based on a certain threshold of use (Erker and Guy, 2012; Bayley et al., 2013; Linford and Shin, 2013; Rodríguez-Ordóñez, 2022). Still, other scholars have defined frequency as a continuous variable, counting the instances of occurrence (Erker and Guy, 2012; Posio, 2015; Purse et al., 2022). Given the inconsistencies that arise in many variationist studies that investigate frequency, it is possible that the asymmetries in defining frequency across studies have made empirical research more challenging.

From a theoretical perspective, questions still remain surrounding the integration of UBG insights into Variationist research. Furthermore, the difficulties in integrating the theoretical aspects of UBG into Variationist Sociolinguistics are amplified when considering the domain of morphosyntax. While previous studies have explored morphological properties in the context of phonetic variation and frequency (Myers and Guy, 1997; Berkenfield, 2001), studies of frequency as a constraint on morphosyntactic variation, per se, are less common in the research literature. With morphosyntactic variation, muscle memory and frequency do not clearly impact morphosyntactic forms as they do phonological forms. High frequency morphosyntactic forms appear to *resist* change or regularization and can even lose their analysability, or their connection to their original, individual components (Bybee, 2010).

While this frequency-motivated resistance effect is borne out in studies of language change (Hooper, 1976; Bybee, 2006), frequency effects on stable variation, or variation that is not leading to language change, are less clear. Erker and Guy (2012) present promising results in their study of Spanish subject pronouns and lexical frequency. Instead of behaving like other social and linguistic constraints, they find that lexical frequency either initiates or enhances the effects of *other linguistic predictors*. Yet, efforts to replicate this finding have uncovered extremely inconsistent results (Bayley

et al., 2013; Linford and Shin, 2013; Posio, 2013, 2015; Bayley et al., 2017; Li and Bayley, 2018; Rivas et al., 2018), leaving our understanding incomplete and leaving many questioning the usefulness of frequency and other UBG insights in Variationist approaches (Poplack, 2001; Bayley et al., 2017). In their study of frequency effects in Spanish SPP variation, Bayley et al. (2017) write “clearly we need additional studies of the role of frequency in SPP variation if we are to understand whether current theorizing about usage based models can be extended to this area of the grammar and to other cases of syntactic variation” (p. 29). This lack of clarity is not limited to research on SPP variation. This is evidenced through (Poplack, 2001). In an effort to test Usage-Based theories in empirical studies of morphosyntax, Poplack’s (2001) Variationist investigation on the irrealis domain in Canadian French presents inconsistent and seemingly unorganized frequency effects across tenses. In other words, the seemingly straightforward UBG predictions linking type (e.g. verb tense categories) and token (e.g. individual verb forms) frequency with form productivity are evidently not borne out in natural speech in Poplack’s study.

Another challenge with incorporating UBG insights into variationist research on morphosyntax is addressing the Usage-Based notion of a prefab, or pre-fabricated construction. Usage-Based grammarians might argue that prefabs become their own separate lexical items with entirely different exemplar clouds –though these clouds are still highly related to the relevant exemplar clouds from which it evolved. If it is the case that prefabs become their own clouds, and all instances of a prefab are invariant, it is potentially in conflict with the sociolinguistic variable, which was defined earlier as “two or more ways of saying the same thing” (Labov, 1972b, p. 188). If a form is prefabricated and therefore can no longer vary, can it reasonably be compared as a variant of a single sociolinguistic variable in variationist research? It is unclear how the UBG prefab and the sociolinguistic variable can work together when they are

seemingly at odds with each other.

In addition to the challenge of wrangling with the UBG notion of a Prefab in Variationist research, some researchers have raised questions surrounding the integration of UBG-inspired constraints (Tamminga et al., 2016). Constraints inspired by UBG are said include frequency metrics as well as aspects of variation and change that appear to be influenced by properties of the human mind, i.e. priming, which include (among other things) a sensitivity to frequency properties of linguistic forms. These constraints are potentially a poor fit in the conventional, dichotomous conception of conditioning factors as social or linguistic. Instead, they are arguably best categorized as a third category separate from social and linguistic factors: ‘cognitive’ factors (though other studies have referred to them as “p-conditioning” Tamminga et al., 2016, p. 2). However, the extent to which these constraints constitute a proper set of constraints such that they are meaningfully different from linguistic and social constraints is still unclear.

One possible reason for the mixed results of previous research on frequency effects in morphosyntactic variation and the lack of clarity on these remaining questions could be Variationists’ tendency to examine frequency in a way that does not capture the entire spirit of exemplar theoretic approaches, i.e. through metrics that take into account (dis)favoring contextual frequency, not just gross frequency of occurrence. This possibility, which will be discussed in more detail in the following section, acts as the primary source of motivation for the present dissertation project, and will assist in illuminating questions (1), (2) and (3). To test this hypothesis, this thesis revisits a promising study of lexical frequency and pronoun use (Erker and Guy, 2012) with a new perspective at hand.

1.3 Spanish Subject Pronoun Use and Usage-Based Grammar

Aptly named a “showcase variable” (Bayley et al., 2012, p. 22), Spanish subject pronominal variation is one of the most well-studied sociolinguistic variables, having been the focus of many Variationist studies (Silva-Corvalán, 1982; Bayley and Pease-Alvarez, 1997; Miyajima, 2000; Flores-Ferrán, 2004; Lapidus and Otheguy, 2005a,b; Otheguy et al., 2007; Erker and Guy, 2012; Vidal Covas, 2013; Carvalho et al., 2015; Bayley et al., 2017; Cacoullos and Travis, 2019; Padilla, 2020 and others). In Spanish, as in multiple other languages (often referred to as *pro-drop languages*), finite verbs variably occur with or without a subject pronoun. Thus, as is shown in Example (1), the subject pronoun can be included (1-a) or excluded (1-b), and the truth conditional properties of the utterances are identical.

- (1) a. Yo quiero caminar
 ‘I want to walk’
- b. Quiero caminar
 ‘I want to walk’

Some researchers have recently attempted to leverage usage-based insights in what are essentially variationist studies of Spanish pronoun use, but the lack of consistent frequency effects in results highlights precisely the unresolved issues motivating the current dissertation (Bayley et al., 2013; Linford and Shin, 2013; Posio, 2013, 2015; Bayley et al., 2017; Rivas et al., 2018). One way that these studies have incorporated some of the insights of UBG is by including a Frequency, or count, of each verb as it appears in their corpora. Adding verb frequency into quantitative models, researchers have attempted to shed light on how frequency impacts variable pronoun use *and* how frequency interacts with other predictor variables that directly impact variable pronoun production. For UBG in the phonological domain, the notion of physical

repetition leads to a relatively straight-forward prediction: forms that are said more frequently will show the reductive effects of streamlined motor routines. With respect to pronoun use, the analogous prediction would favor reduced utterances. That is to say that insofar as the non-use of a pronoun is a more reduced utterance, then high frequency verbs might be expected to occur with pronouns at overall lower rates. However, the link between language use and the physical act of speech is less clear outside of phonology, since previous research on morphosyntax has shown the opposite effect of frequency on pronoun use. Erker and Guy (2012) found an extremely weak but nonetheless significant correlation between lexical frequency and subject pronoun variation that showed higher frequency corresponding to increased pronoun use. They also found that frequency modulated the effects of all other linguistic predictors. Specifically, high lexical frequency either amplified an existing effect of other, previously established linguistic predictors, or it activated an effect of other linguistic constraints that was only present in high frequency forms. Due to an inconsistent main effect of frequency, high frequency does not seemingly correlate with either systematically lower or higher rates of pronoun use. Instead, Erker and Guy (2012) conclude that high frequency corresponds to increased variation in pronoun rates among the most frequent forms in their study. This finding is incongruous with how variationist conditioning factors are thought to work: typically, conditioning factors act as constraints on the sociolinguistic variable in question, not as a constraint on *all* of the other conditioning factors, and they generally do so with consistency across forms (although some lexical exceptions do exist). Moreover, the findings in Erker and Guy (2012) are somewhat unique, since other attempts to replicate this finding in studies on pronoun variation have come up short, finding amplification effects on a small subset of constraints (Linford and Shin, 2013; Posio, 2015), or finding none at all (Bayley et al., 2013; Posio, 2013; Bayley et al., 2017; Rivas et al.,

2018; Lease et al., 2022).

From a purely statistical perspective, incorporating lexical frequency into variationist research is an obvious starting point, since frequency metrics are easily integrated into conventional variationist statistical models (i.e. in model construction). Nevertheless, studies on morphological variation that have investigated lexical frequency have often done so in a manner that arguably undertheorizes Usage-Based Frameworks, separating frequency from the multi-faceted structure of the exemplar clouds posited to form the cognitive map in UBG. That is to say that previous morphosyntactic variationist research that has attempted to understand the role of Frequency has not endeavored to do so in a manner that maintains the robust theory of UBG, which postulates detailed remembered experiences organized based on many characteristics across many levels of linguistic structure. An absolute frequency metric does not consider the detailed linguistic and social information that speakers are storing in their mental grammars, since counting instances of use does not include the contexts in which verbs are used.

In light of the issues raised above – general challenges of integration of UBG insights into Variationism, and the lingering uncertainty associated with the role of frequency in one of the most widely studied (non-phonological) variable phenomena – I propose to do the following: (1) replicate Erker and Guy (2012) because of the frequency effects they found which Bayley et al. (2017) and others could not replicate (Bayley et al., 2013; Posio, 2013; Bayley et al., 2017; Rivas et al., 2018; Lease et al., 2022); (2) investigate the constraining potential of multiple novel usage-based metrics (Chapter 5), in an effort to more robustly integrate aspects of UBG into Variationist research.

In the replication, which is described in Section 2.2, quantitative analyses (e.g. mixed-effects regression modeling) are used to investigate the conditioning potential

of a series of linguistic variables that have been known to impact Spanish pronoun use. This portion acts as a baseline (Chapter 4), and includes the frequency metrics used in Erker and Guy (2012): Raw Frequency, Log Frequency, and Discrete Frequency. Each of these frequency metrics is based on the number of occurrences of the potentially-pronoun-hosting finite verb in the corpus.

For the expanded usage-based measures, the present dissertation probes a series of contextual frequency metrics that consider the rate at which the verb appears in specific linguistic contexts. Measuring frequency based on the linguistic context is in line with the UBG notion that speakers are aware of and retain great detail with each linguistic experience. Contextual frequencies, like those proposed here, represent a significant, if still modest, increase in fidelity to UBG conceptions of linguistic knowledge. Since exemplar theoretic models of language posit that exemplar clouds store every variation of a form, investigating a contextual frequency, which considers different contexts, incorporates some of this variation into empirical research on spoken language. Crucially, this kind of variation allows the current dissertation to compare not only frequencies across lexical items, but also frequencies within lexical items. Pronoun rates can be compared for all finite verbs that possess the same contextual frequency, but pronoun rates can also be compared for a single finite verb based on the number of occurrences in different levels of the contextual frequency. Consider a contextual frequency that counts the frequency at which a verb appears with a switch in referent from the previous verb. In addition to comparing the pronoun rates across forms for a given frequency in a switch reference context, contextual frequency metrics also allow for a comparison of rates between contexts for a single form.

Finally, the dissertation wrangles with the joint findings from the replication and expansion, ultimately shedding light on the theoretical implications of incorporating

UBG insights into Variationist research. Data comes from the Spanish in Boston Corpus (Boston) and the Spanish in New York Corpus³ (NYC), making this replication magnitudes larger. These corpora consist of sociolinguistic interviews conducted with Spanish speakers living in either Boston, MA or New York, NY ($N_{speakers} = 80$ & 141 respectively). Taken together, the present dissertation is the largest study of variable Spanish pronoun production to date.

1.4 Research Questions

The specific research questions that the present dissertation will answer are outlined below:

1. Does replication with a larger, different data set produce results in line with those of Erker and Guy (2012)?
2. Does incorporating contextual frequency metrics provide clearer results?
3. What are the theoretical implications of incorporating insights from Usage-Based Grammar into Variationist Sociolinguistic research on Spanish subject pronoun production?
4. To what extent can the study's results shed light on the notion of a prefab?

1.5 Broad Predictions and Overview of the Results

Several studies have attempted to replicate the frequency effects found in Erker and Guy (2012) (in Spanish: Bayley et al., 2013; Linford and Shin, 2013; Posio, 2013, 2015; Bayley et al., 2017; Rivas et al., 2018 and in Mandarin: Li and Bayley, 2018).

³It is important to note that this study would not be possible without the hard work and generosity of Ricardo Otheguy, Ana Celia Zentella, and Daniel Erker. Many thanks to them and all of the researchers that had a hand in creating these corpora.

Some studies have found no effect of lexical frequency (Posio, 2013; Rivas et al., 2018). While others have found that high frequency disfavors pronoun use, which contrasts the results in Erker and Guy (2012) (Bayley et al., 2013, 2017). Nevertheless, a few studies have found a mediating effect of lexical frequency, albeit weak and inconsistent (Linford and Shin, 2013; Posio, 2015).

This mix of findings from previous studies create some challenges in formulating confident predictions. Furthermore, none of these studies have investigated the role of verb frequency in a corpus of this magnitude. The original Erker and Guy (2012) study investigated speech from 12 speakers. Bayley et al. (2013) examined the speech of 29 speakers. Linford & Shin's (2013) study had 12 speakers. Rivas et al. (2018) examine the pronoun use of 32 speakers. Since the number of participants in the current dissertation is seven times larger (221 speakers) than the largest previous study of this kind, I predict that a larger, different dataset will provide more fine-grained results that fall in line with Erker and Guy (2012). Additionally, the context-dependent frequency metrics presented in the current study are hypothesized to paint an even clearer picture of the extent to which speakers are sensitive to verb context. That is to say that the verb frequencies studied here, which contain additional detail, are predicted to significantly impact subject pronoun production. This, in turn, would support the UBG notion that speakers are acutely aware of and sensitive to detailed information regarding verbs' contexts and how these contexts relate to pronoun presence/absence.

Results from the present dissertation show strong support for Erker and Guy's (2012) results: Discrete Frequency is the most explanatory lexical frequency metric, and it acts as an amplifier of the effects other linguistic constraints on pronoun use. Replication results also reveal a statistically significant, but opaque and inconsistent, main effect of lexical frequency on pronoun production. The investigation of con-

textual frequency metrics proved to be much more challenging. Due to the overall frequency effect reported in the replication, the contextual frequency effects were only transparent in a subset of the contextual frequency metrics. These revealed that contextual frequency does impact pronoun use as expected, but only for verb forms that occur at high-enough overall frequencies. Nevertheless, results indicate that contextual frequency metrics do a better job of accounting for pronominal variation than simply overall lexical frequency. These results support UBG notions that speakers are sensitive to details surrounding the contextual properties of linguistic forms. However, sensitivity to the contextual tendencies of a given form does not appear until that form has built up a large enough mental representation, indicating that lexical frequency and contextual frequency impact variation.

1.6 Dissertation Structure

The current project consists of six chapters: (1) Introduction, (2) Background, (3) Methodology, (4) Replication Results, (5) Expansion Results and (6) Conclusions. The present chapter introduces the topic, discusses the motivations for the research, and outlines the structure of the dissertation. The Background chapter recounts previous research relevant to the proposed study on Spanish SPPs, specifically in the fields of Variationist Sociolinguistics and Usage-Based Theories of Language. In addition to developing the foundation upon which I build my dissertation, Chapter 2 also highlights the areas of open inquiry, especially as they pertain to previous research that has included insights across UBG and Variationist Sociolinguistics. In Chapter 3, I outline the methodology for the study. The study incorporates the frequency metrics outlined in Erker and Guy (2012), and also investigates the impact of Contextual Frequency Metrics, offering a novel set of variables that better retain the spirit of UBG. Chapter 4 presents the results that are a direct replication of

Erker and Guy (2012), which act as a baseline to inform the expanded frequency results in Chapter 5. In Chapter 5, I report the results of the novel contextual frequency metrics. Finally, Chapter 6 concludes the dissertation with a discussion of the findings, a summary of the major takeaways, and a description of some remaining questions or concerns.

Chapter 2

Review of Existing Literature

The research questions in the present dissertation draw from: (1) Variationist Sociolinguistic research, (2) research on Spanish Subject Pronoun variation, and (3) Usage Based Theories of Language. The current chapter provides an overview of each of these lines of inquiry, describes their research history as it relates to investigating frequency effects, and outlines in more detail the remaining questions that inform the current dissertation, which were sketched in the previous chapter. First, we begin with the Variationist Sociolinguistic research, then move to Spanish Subject Pronoun variation more specifically. Finally, we move to a discussion of UBG research, before presenting the specific unanswered questions that act as the motivations for the current study in greater depth.

2.1 Variationist Sociolinguistics

Variationist Sociolinguistics is primarily interested in linguistic variation as a window into society and how social structures relate to language use, which in turn influences language change. Variationist Sociolinguistics considers variation an inherent property of human language, and as such, a central characteristic of linguistic knowledge. Investigations on the role of lexical frequency in variationist research extend back multiple decades. This is especially true in the context of phonetic variation and sound change (Myers and Guy, 1997; Berkenfield, 2001; Brown and Torres Cacoullos, 2003; Labov, 2006; Abramowicz, 2007; Díaz-Campos and Gradoville, 2011;

Brown and Raymond, 2012; Raymond and Brown, 2012; Bongiovanni, 2014; Tamminga, 2014; Brown, 2015; Hay et al., 2015; Escalante, 2016; Brown et al., 2021; Purse et al., 2022). Chapter 1 briefly describes one of the earliest variationist studies to consider lexical frequency (Myers and Guy, 1997). Their study on coronal stop deletion (CSD) suggests that frequency affects phonetic reduction phenomena differently depending on the word category: high-frequency monomorphemic words (e.g. *lift*) favor CSD, but past tense forms (e.g. *laughed*) show no significant difference in deletion rates across frequencies. Berkenfield (2001) finds that the frequency of “functional categories” (e.g. demonstrative adjectives or complementizers) impacts vowel duration of English *that*. In other words, as the use of *that* in specific functional categories increases, the vowel in *that* decreases in duration. These findings suggest that lexical items with different functional categories may be stored as separate lexical entries.

In more recent work on CSD, Purse et al. (2022) investigate three different lexical frequency metrics: whole word frequency (the full form of a word), stem frequency (the lemma of a word, e.g. *likes*, and *liked* have the same lemma *like*), and conditional frequency (the relative frequency of a form based on whole-word frequency divided by lemma frequency). They use these metrics to assess the relationship between the mental grammar and actual language use, and to uncover whether all lexical frequencies are “created equal” across variable phenomena. Their results indicate that conditional frequency has the strongest, and most consistent, direct impact on CSD for monomorphemic words and past tense *-ed* words. Much like Erker and Guy (2012), they also find an amplification effect of frequency. Specifically, their results indicate that whole-word frequency amplifies other constraints: CSD is significantly higher in frequent monomorphemic forms than in regular past tense forms, but there is virtually no difference in CSD for infrequent monomorphemic or past tense forms.

Díaz-Campos and Gradoville (2011) investigate lexical frequency effects on Spanish intervocalic /d/ deletion by Colombian speakers and variable /ʒ/ devoicing by Argentinian speakers. They analyze two different lexical frequency metrics: word frequency and type frequency. They find that lexical frequency and type frequency impact /d/ deletion in the direction that is predicted by UBG: high frequency forms and types favor /d/ deletion compared to low frequency forms/types. For /ʒ/ devoicing, they find that forms that occur in devoicing-disfavoring contexts (e.g. *yo* ‘I’ appearing after a pause) more frequently impact overall devoicing for that form. This study provides strong support for the UBG theory surrounding the non-uniform nature of linguistic variation and language change: the more frequently that forms appear in change-favoring contexts, the further along they are in their change than forms that appear less frequently in change-favoring contexts.

In recent years, some sociolinguistic research, particularly that concerned with child or adult acquisition of sociolinguistic variation, have investigated the role of lexical frequency on variation (Shin, 2016; Kanwit and Geeslin, 2020; Callen and Miller, 2022; Rodríguez-Ordóñez, 2022; Lease et al., 2022). These studies have set out to better understand the extent to which frequency of exposure to linguistic forms impacts the acquisition of certain sociolinguistic variables (i.e. mastering the use of the variants of a sociolinguistic variable) or the sensitivity to relevant conditioning constraints. Generally, these studies find that the acquisition of sociolinguistic variables and their sensitivity to relevant conditioning constraints emerges over time based on the frequency of forms and the salience of their constraints. In their study on L1 acquisition of Spanish pronominal variation, Shin (2016) aims to understand how morphosyntactic variation, and constraints on that variation, are acquired by children. Results from this study show that Mexican children in the study are sensitive to certain conditioning factors (first person singular such as ‘I walk’ and a switch in

referent from the previous finite verb) regardless of age-group (6–16 years old). However, their findings also show that other constraints (Tense-Mood-Aspect of the verb and semantic verb category) are learned later in adolescence. These findings suggest that children acquire patterns of structured variation more quickly when they are relevant to frequent forms, and as children gain linguistic input through experiences over time, sensitivities to constraints on less-frequent forms accumulate. Kanwit and Geeslin (2020) examine the extent to which adult L2 learners of Spanish acquire copula variation and sensitivity to the constraining variables that impact copula variation. Overall, their findings show that advanced learners of Spanish approximate native speaker treatments of constraints more than less advanced Spanish learners. These results indicate that type and token frequency impact acquisition of variable forms.

Rodríguez-Ordóñez (2022) investigates the role of frequency on the acquisition and use of ergative case marking for L2 learners of Basque. In their study, they compare ergative case usage by L2 Basque learners, fluent L2 Basque speakers, and L1 Basque speakers. Overall, they find that frequency is not a significant independent factor for native speakers. However, when they split their data into two groups based on discrete verb frequency, they find a frequency effect consistent with Erker and Guy (2012). The infrequent verb model shows minimal significance of linguistic predictors on ergative case marking for fluent and new L2 Basque speakers. In contrast, all linguistic predictors are found to be significant for frequent verbs for fluent L2 speakers, and new L2 speakers show increased sensitivity to linguistic predictors for frequent verbs. These results are encouraging for UBG, as they point to the cumulative effect of experiencing linguistic forms. Over the course of L2 acquisition, as a learner accumulates experiences with a linguistic form, they also accumulate a sensitivity to the linguistic and social constraints that influence that form. It is in-

interesting that Rodríguez-Ordóñez (2022) do not find an effect of frequency for native speakers, since Erker and Guy (2012) were analyzing the speech of native Spanish speakers. However, this difference could suggest that frequency effects are not consistent cross-linguistically or that they impact different sociolinguistic variables in different ways. This potential variation in frequency effects is not unfounded: other, more well-understood sociolinguistic variables are constrained by different features, and these features occur at different rates. The constraints that impact variable ergative case marking in Basque are ostensibly different from the constraints that impact variable pronoun use in Spanish, and these differences likely impact frequency effects. As for cross-linguistic differences, although they are beyond the reach of the current dissertation, studies have shown that speakers of different languages produce semantically similar forms (or translational equivalents) at different frequencies (Dionne and Coppock, 2022).

Perhaps most relevant to the research questions addressed in this dissertation, Raymond and Brown (2012); Brown (2015); Raymond et al. (2016); Brown et al. (2021) have all researched the role of context-based frequency metrics on various aspects of sociolinguistic variation. Brown (2015) investigates the relationship between context-dependent frequency and word-initial /d/ reduction in Spanish. For their variable, the context-dependent frequency is defined as the proportion at which each word appears after all sounds except for nasals, laterals, and pauses (only preceding nasals, laterals and pauses favor /d/ production). They find a significant effect of the reduction-favoring contextual frequency, such that as reduction-favoring frequency increases, the likelihood of /d/ reduction increases. They are also able to tease apart cognate effects (i.e. that English cognates' /d/-reduction patterns differently), showing that these effects are actually due to contextual frequency effects. Ultimately, this study provides strong support for the cumulative impact of reduction contexts:

the propensity for a form to appear in a reduction favoring context increased the likelihood of reduction over all contexts. Also concerned with variable sociophonetic reduction, Raymond et al. (2016) research the impact of cumulative exposure of lexical forms to reduction-favoring environments on word-initial /s/ reduction in Spanish. In their study, which examines /s/ in the speech of New Mexican Spanish speakers, they find that how often a word appears in a reduction-favoring environment (proportion of observations that follow a mid or low vowel), and not how often the word appears overall, significantly impacts word-initial /s/ reduction. They assert that this result supports a change in mental representation for higher-frequency forms, such that exemplar clouds for higher frequency forms accumulate information about the phonological contexts in which they most often appear. These studies present important evidence that supports usage-based grammar theories of language use, and show that lenition phenomena aren't just more likely in frequent forms, but they are more likely in forms that frequently occur in reduction-favoring contexts. As with other variationist work on frequency effects, these studies of contextual frequency have their fullest and most persuasive articulation in phonetic variation. As we will see in Section 2.2.3, more recent work has leveraged the contextual frequencies investigated in phonetic variation and has applied similar methods to investigate morphosyntactic variation (Brown and Shin, 2022; Lease et al., 2022).

2.2 Variationist Sociolinguistic Research on Spanish Subject Pronouns

As mentioned in Chapter 1, variable Spanish subject pronoun production is one of the most well-documented, and arguably most well-understood, variables in the field. Variationist studies on this feature have shed light on the linguistic and social predictor variables that influence pronoun use. The robust nature of this specific line of

research cannot be overstated. The following subsections outline a series of the most relevant linguistic and social predictor variables that have been shown to significantly constrain pronoun production for Spanish speakers.

2.2.1 Linguistic Constraints on Variable Pronoun Production

Multiple linguistic predictor variables have been investigated as potential constraints on Spanish subject pronoun use. To illustrate some of these variables and their different levels, let us consider the excerpt in Example (1), taken from the Otheguy Zentella Corpus of Spanish in New York (Otheguy and Zentella, 2012):

- (1) N: Cuando (38) yo voy llegando por Indubán, que (39) ∅ voy entrando ya para mi casa, (40) ∅ voy caminando, y (41) yo pendiente de que iba a llegar a mi casa, so [pronuncia en inglés] cuando (42) yo veo que (43) ∅ veo un chamaquito en una.. en una.. en una.. bicicleta, que (44) ∅ sube de repente así a la acera, y (45) ∅ me hace así, y (46) ∅ me hace así en la cadena y (47) yo nada más [namás] siento cuando (49) él me toca, [48 skipped by transcriber], (50) ∅ óyeme, (51) ∅ mira, eso (52) ∅ fue un susto que (53) yo me quedé así como en shock, que (54) yo no sabía lo que (55) yo iba a hacer, pero lo curioso del asunto (56) ∅ fue que la cadena.. no (57) ∅ sé si la medalla (58) ∅ se cayó en el suelo o (60) él se la cogió, no (61) ∅ sé porque (62) ∅ fue tan rápido que lo (63) ∅ hizo, que el shock (64) ∅ fue tan grande, que la cadena, cuando (65) yo llegué a mi casa, que (66) ∅ me cambié y que (67) ∅ me quité el XXX que (68) yo tenía, ahí (68a) ∅ estaba la cadena, rota, (68b) ∅ se quedó ahí, y (68c) yo no sé como no.. porque un tramo de dos esquinas[equina] que (68d) yo caminé para llegar a mi casa, (68e) yo no sé como esa cadena no se (68f) ∅ me perdió.

Gloss: When (38) I am arriving through Indubán, that (39) \emptyset am almost entering my house, (40) \emptyset am walking, and (41) I was focused about arriving to my house, so when (42) \emptyset [I] see that (43) \emptyset [I] see a little boy on a... on a... on a... bicycle, who (44) suddenly climbs onto the sidewalk like this, and (45) \emptyset does this to me, and (46) \emptyset does this on the chain and (47) I just (just) feel when (49) he touches me, (48 skipped by transcriber), (50) \emptyset listen to me, (51) \emptyset look, that (52) \emptyset was a scare that (53) I was like in shock, that (54) I didn't know what (55) I was going to do, but the curious thing about it (56) \emptyset was that the chain... (57) \emptyset [I] don't know if the medal (58) \emptyset fell on the ground or (60) he caught it, (61) \emptyset [I] don't know why (62) \emptyset was so fast that (63) \emptyset did it, that the shock (64) \emptyset was so great, that the chain, when (65) I got home, that (66) \emptyset changed and (67) \emptyset took off the XXX that (68) I had, there (68a) \emptyset was the chain, broken, (68b) \emptyset it stayed there, and (68c) I don't know how it didn't.. because a section of two corners that (68d) I walked to get to my house, (68e) I don't know how that chain didn't (68f) \emptyset [I] lost it

Morphological Regularity

One linguistic variable that has been shown to impact subject pronominal variation is “Morphological Regularity” (Erker and Guy, 2012; Bouchard, 2018). Studies suggest that finite verb forms that are regular relative to their underlying infinitival forms are more likely to occur with overt pronouns than irregular verb forms. This effect is hypothesized to occur due to the tendency for irregular verb forms to have very unique forms for various person/number levels. A functionalist account (first set forth by Hochberg, 1986) argues that when forms are unique, they are less likely to be ambiguous and can therefore occur without a pronoun (Otheguy and Zentella, 2012). One example of an irregular verb form is illustrated in sentence (38) *yo voy llegando* ‘I am arriving’. *veo* ‘I see’ in sentence (42) is an example of a regular verb form.

Person and Number

The person and number properties of verbs and subjects (referred to as person/number moving forward) has also been shown to impact pronoun production in Spanish (Bayley and Pease-Alvarez, 1997; Lapidus and Otheguy, 2005a; Otheguy et al., 2007; Torres Cacoullos and Travis, 2011; Anderson, 2013; Gudmestad et al., 2013; Alfaraz,

2015; Orozco, 2015, 2018; Orozco and Hurtado, 2021; Padilla, 2021). Specifically, research suggests that singular subjects favor pronouns more than plural subjects. Additionally, first and second person singular forms have been shown in some data to favor pronouns more than third person singular forms. It is hypothesized that speakers prefer pronouns in singular, first, and second person forms because the conjugations of finite verbs have increased tendency towards ambiguity for the listener (Otheguy and Zentella, 2012). Therefore, the inclusion of an SPP removes potential confusion for the listener. The subject pronoun in sentence (38) is in first person singular, whereas the pronoun in (49) is in third person singular (*yo* ‘I’ and *él* ‘he’, respectively). Other things being equal, the literature leads us to expect that sentence (38) would occur with a pronoun, and sentence (49) would not.

Tense, Mood, and Aspect

The tense, mood, and aspect properties of verbs (TMA) have also been reported to impact variable pronoun use (Silva-Corvalán, 1982; Hochberg, 1986; Bayley and Pease-Alvarez, 1997; Flores-Ferrán, 2002, 2004; Otheguy et al., 2007; Travis, 2007; Prada Pérez, 2009; Erker and Guy, 2012; Shin, 2014; Torres Cacoullos and Travis, 2015). Studies suggest that preterite, perfective, future, and imperative forms correlate to lower overt pronoun production than forms such as the imperfect indicative, and conditional forms. This is hypothesized to occur due to the tendency for imperfect indicative and conditional forms to be identical across multiple person/number combinations. For instance, imperfect indicative first- and third- person singular forms are identical for regular verbs, as well as second person singular and plural. Sentence (43) is in present tense, while (66) in Example (1) (*∅ me cambié* ‘I changed’) is in preterite tense.

Semantic Category

Historically, research has consistently supported the claim that the semantic category of the verb affects pronoun production (Bentivoglio, 1980; Enriquez, 1984; Flores-Ferrán, 2002, 2004; Otheguy et al., 2007; Orozco and Guy, 2008; Abreu, 2009; Carvalho and Child, 2011; Posio, 2011; Otheguy and Zentella, 2012; Erker and Guy, 2012; Torres Cacoullos and Travis, 2018). Specifically, research has shown that mental activity verbs (e.g. *pensar* ‘to think’) correspond to the highest overt pronoun rates, followed by stative verbs (e.g. *estar* ‘to be’), then external activity verbs (e.g. *comprar* ‘to buy’). While these three factor values are common, and are used in Erker and Guy (2012), other methods for defining Semantic Category have been used as well. In the first study of semantic category on variable pronoun use, Bentivoglio (1980) investigated five levels: (1) cognitive verbs (e.g. *pensar* ‘to think’), (2) perceptive verbs (e.g. *ver* ‘to see’), (3) enunciative verbs (e.g. *decir* ‘to say/tell’), (4) desiderative and manipulative verbs (e.g. *ordenar* ‘to command’), and (5) other verbs. Since then, other studies have investigated four semantic categories, separating “estimative” verb forms (e.g. *admirar* ‘to admire’) out of the mental activity class (Enriquez, 1984; Flores-Ferrán, 2002, 2004; Carvalho and Child, 2011). Sentence (49) in Example (1) (*él me toca* ‘He touches me’) contains the external activity verb *toca*. *Sabía* ‘knew’ in sentence (54) (*yo no sabía* ‘I didn’t know’) is an example of a mental activity verb. The predicted differential pronominal tendencies are that the external activity verb in (49) disfavors a pronoun while the mental activity verb in (54) favors a pronoun.

Recently, studies have called the effect of semantic class into question (Orozco and Hurtado, 2021; Travis and Torres Cacoullos, 2021, Orozco, 2022). Orozco and Hurtado (2021) present data that suggest that the effects of semantic categories are actually the underlying impact of the pronominal tendencies of a few high frequency verb forms. In contrast to an effect of semantic class, they find that a few high frequency verb

forms are skewing the data in the direction of the effect that is seemingly observed for semantic category. That is to say that the significant differences in pronoun rates for verbs in different semantic categories are actually caused by individual verbs in each category that are extremely high in frequency relative to the other forms in the category. For example, *creer* ‘to believe’ and *pensar* ‘to think’, which had pronoun rates of 88.2% and 74.7% respectively in Orozco and Hurtado’s data, likely skew the pronoun rate for mental activity verbs.

Switch Reference

Research shows that Spanish speakers are more likely to use a subject pronoun in contexts without subject continuity, i.e. with a switch in reference (Bentivoglio, 1987; Cameron, 1992; Orozco and Guy, 2008; Otheguy and Zentella, 2012; Erker and Guy, 2012; Alfaraz, 2015; Carvalho and Bessett, 2015; Michnowicz, 2015; Bessett, 2018; Padilla, 2021). The hypothesized reason for this is to avoid confusion or ambiguity for both the speaker and the listener: when a speaker begins talking about a new referent in dialogue, signaling this change can ease processing load for both (Otheguy and Zentella, 2012). There have been multiple methods for defining switch reference, but the definition put forth in the current dissertation (and taken from Erker and Guy (2012) is the following: whether the grammatical subject is different or the same as the grammatical subject in the immediately preceding verb phrase. I will refer to this as “Switch Reference”. Sentence (44) represents a site where there is a switch in reference since the previous grammatical subject, in (43) is *yo* ‘I’ and the grammatical subject in (44) is *él* ‘he’. Switch reference is most often defined as a binary variable that considers the immediately preceding grammatical subject. However, some variationist research has operationalized switch reference as a quasi-continuous variable based on the relative distance from the referent (Givón, 1983; Myhill, 2005). Further, other studies have narrowed the categorical variable to only

consider switches in human referents (Travis and Torres Cacoullos, 2012). Travis and Torres Cacoullos (2012) find an effect of interfering human referent that is stronger than the effect of an intruding nonhuman referent.

Preceding Pronoun

Spanish speakers are also more likely to use a pronoun when the previous site of pronominal variation contains a subject pronoun, and they are less likely to use a pronoun when the previous site of pronominal variation does not contain a pronoun. The underlying logic here is that the online recency effect of a previously produced or not produced pronoun lingers for speakers. This influence of previous pronoun presence or absence is often referred to as “Priming” (Cameron, 1992; Flores-Ferrán, 2002; Cameron and Flores-Ferrán, 2004; Travis, 2007; Abreu, 2009; Torres Cacoullos and Travis, 2011; Carvalho and Child, 2011; Abreu, 2012; Shin, 2014), but it will be called Preceding Pronoun in the current thesis. It has been operationalized as a binary present/absent, but it has also been operationalized as a continuous variable that counts the number of intervening clauses between a token and the previous pronoun. As we see in Example (1), sentence (54), which contains an overt subject pronoun, is purportedly primed for pronoun presence (rather than absence) by the overt subject pronoun in the previous grammatical subject (53).

2.2.2 Social factors and Their Impact on Spanish Subject Pronouns

In addition to many linguistic variables that have been investigated alongside Spanish subject pronoun production, a series of social variables have been studied. Some of these variables include age, sex, region of origin, level of education. Overall, there is a general trend for studies on subject pronoun variation that suggests that external, social constraints do not significantly impact subject pronoun production. Region of Origin has shown the most consistently significant results, with Caribbean and coastal

South American Spanish speakers favoring overt pronouns more than other regions (such as Andean, Central American, or Peninsular varieties) (Orozco and Guy, 2008; Otheguy and Zentella, 2012; Carvalho et al., 2015).

Age, gender, and education level have shown consistently mixed results, with some studies finding significant effects for these predictors and others finding the opposite or no effect (Cameron, 1992; Ávila-Jiménez, 1995; Bayley and Pease-Alvarez, 1997; Flores-Ferrán, 2002; Otheguy et al., 2007; Orozco and Guy, 2008; Carvalho and Child, 2011; Otheguy and Zentella, 2012; Holmquist, 2012; Shin, 2013; Shin and Otheguy, 2013; Prada Pérez, 2015; Lastra and Butragueno, 2015; Michnowicz, 2015; Lapidus Shin and Erker, 2015). Many studies have reported no effect of age in their investigation of Spanish subject pronoun use (Cameron (1992); Michnowicz (2015) find no age effects, while Ávila-Jiménez (1995); Flores-Ferrán (2002) suggest that younger speakers favor pronoun use and Orozco and Guy (2008); Lastra and Butragueno (2015) report that older speakers favor pronoun use. Similarly, studies have shown conflicting effects for gender: Otheguy et al. (2007); Orozco and Guy (2008); Holmquist (2012) report no effect of gender, while other studies suggest women favor pronouns more than men (Bayley and Pease-Alvarez, 1997; Carvalho and Child, 2011; Otheguy and Zentella, 2012; Shin, 2013; Shin and Otheguy, 2013; Lapidus Shin and Erker, 2015). For education, research has been particularly sparse: Lastra and Butragueno (2015) find that speakers with less education produce higher rates of SPPs than speakers with more education, though this result was not statistically significant.

Some studies have investigated the role of immigration history and/or contact-induced change on pronoun use by Spanish speakers in the U.S. (Otheguy and Zentella, 2012; Torres Cacoullos and Travis, 2015, 2018; Erker, 2022). One way that some studies have operationalized immigration (which is used as a measure for language

contact) is through Percent of Life in the U.S. (PLUS). Research has shown a significant effect of PLUS, such that increased PLUS corresponds to increased pronoun use (Erker, 2022). In their seminal study of Spanish in New York City, Otheguy and Zentella (2012) find a similar effect of immigration. Specifically, non-Caribbean established immigrants (arriving in NYC before 17 and living there for more than 5 years) produced higher pronoun rates than the non-Caribbean newcomers (arriving in NYC after 17 and living there for less than 5 years). They argue that this finding suggests evidence of the effects of language and dialectal contact: non-Caribbean established immigrants have been in contact with Spanish speakers of Caribbean origin and English speakers (two groups that generally favor pronoun production) for a longer period of time. Nevertheless, the question of whether language contact (particularly Spanish in contact with English) is contributing to language change with respect to Spanish pronoun use is still unclear. Research has shown some evidence of contact-induced change while other research has found no such evidence. Torres Cacoullós and Travis (2018) investigate this very question within a community of speakers in New Mexico. They generally find no difference in pronoun use or sensitivity to constraints between generations, asserting that no change is occurring.

These studies highlight the gaps in knowledge surrounding the impact of social factors on variable pronoun production. While it is clear that further exploration of these factors is in order, the investigation of social constraints is outside of the scope of the current dissertation. The current dissertation, as it stands, exploits the best possible point of overlap between a variationist and exemplar theoretic approach to subject pronoun variation. The social signaling potential of subject pronouns is quite low compared to that of other variable phenomena. In other words, examining the influence of linguistic and discursive factors, while simultaneously setting aside social factors, is the most straightforward method for investigating this morphosyntactic

variable in this manner. Nevertheless, understanding the nature of this particular area of research is relevant for the discussion portion of the dissertation. The next section reports the research history of Spanish subject pronoun variation as it relates to Frequency.

2.2.3 The Role of Frequency in Spanish Subject Pronouns

More recent work on Spanish subject pronoun use has attempted to test hypotheses based on usage-based theories in their sociolinguistic analysis through the addition of verb frequency as a predictor variable (Erker and Guy, 2012; Bayley et al., 2013; Linford and Shin, 2013; Bayley et al., 2017; Rivas et al., 2018; Brown and Shin, 2022; Lease et al., 2022). Erker and Guy (2012), which was described in some detail in Chapter 1, analyzed a series of linguistic predictor variables in tandem with a variable that measured verb frequency within their corpus. They find a weak, but significant main effect for raw verb frequency on pronoun use such that increased frequency corresponds to a slight increase in pronoun use. Log frequency of the verb was significant, albeit weak, showing the opposite effect: increased log frequency corresponded to decreased pronoun use. They also found that the categorically coded lexical frequency of the verb (i.e., coded as either “frequent” or “infrequent” based on a relatively arbitrarily decided threshold of 1%) was a statistically significant predictor for the response variable on its own. Verbs categorized as “frequent” showed higher pronoun rates, but “infrequent” verbs corresponded to lower pronoun rates. Although main effect results were statistically significant, they observe that SPP rates are poorly predicted by frequency for a large fraction of verb forms, and the predictions do not improve with increasing frequency. This leads them to conclude that frequency is a poor predictor of pronoun use on its own.

When incorporating other independent variables, however, Erker and Guy (2012) find that lexical frequency interacts with all other predictor variables – increasing,

and in some cases activating, their effects for high frequency verbs and decreasing their effects when considering low frequency verbs. Their results suggest that variant choice is fundamentally related to the frequency properties of verbs. That is to say that speaker experiences impact the extent to which certain linguistic factors influence variant choice. As mentioned in Section 1.5, replicating the quantitative effect found in Erker and Guy (2012) has been challenging (Bayley et al., 2013; Linford and Shin, 2013; Posio, 2013, 2015; Bayley et al., 2017; Li and Bayley, 2018; Rivas et al., 2018), leaving the relevance of frequency as a potentiator inconclusive. In the first replication of Erker and Guy (2012), Bayley et al. (2013) investigate the impact of verb frequency on pronoun production and a series of other linguistic predictors. Their findings differ from Erker and Guy (2012): they find a significant effect of frequency such that frequent verbs *disfavor* pronoun use, and infrequent verbs favor pronoun use. They do not find any amplifying or activating effects on other constraints (with the exception of semantic category amplified a bit for frequent forms).

Linford and Shin (2013) explore frequency effects on pronoun use by early and late L2 learners of Spanish. They found that frequency significantly impacted pronoun use for early learners, but not for late learners. However, they did find that frequency mediated the effects of two of their linguistic predictors (Semantic category and TMA) and not the others (switch reference and person). Their results provide some support for Erker and Guy (2012). Posio (2013) conducts a study on pronoun use by Peninsular Spanish speakers and European Portuguese speakers. They investigate lexical frequency effects as they pertain to the formulaic nature of highly-frequent finite verbs and their impact on pronoun use cross-linguistically. In contrast to the frequent forms in Erker and Guy (2012), extremely highly frequent forms in their corpus were *not* sensitive to other linguistic predictors, and instead behaved in a formulaic manner. In a follow-up study, Posio (2015) finds further support for the

formulaic nature of high-frequency forms. In a corpus study of Peninsular Spanish, Posio (2015) concludes that there is variation in the discourse contexts in which high frequency finite verbs occur, which, in turn, motivates the variation in pronoun rates for different high-frequency verbs. Nevertheless, in this study, they find some support for Erker and Guy (2012), due to uncovering amplification effects of frequency in their qualitative analysis of different semantic categories of finite verbs. Rivas et al. (2018) carry out the first study of SPP variation that explores pronoun use with finite *and* infinitival verb forms (through the construction [*para* + SUBJECT + INFINITIVE]). Results show that, unlike in Erker and Guy (2012), even the extremely infrequent infinitival construction shares identical sensitivities to other linguistic constraints as more frequent finite forms.

Travis and Torres Cacoullos (2021) investigate the role of frequency in cognition verbs and their pronominal tendencies. They aim to understand the extent to which cognition verbs (e.g. *creo*, *pienso* ‘I believe’, ‘I think’) are a true semantic verb category that impacts subject pronominal production, or if the effects of this category are actually the effects of individual high-frequency forms. Their findings suggest that cognition verbs do form a cohesive semantic verb class, but verb classes in general are structured around the tendencies of a few extremely high-frequency forms. Cognition verbs are overwhelmingly produced in first person singular, and the high frequency verb forms at the center of this category seem to be “lexically particular constructions” (Travis and Torres Cacoullos, 2021, p. 3). This suggests that the cognition verb class is actually skewed due to the presence of constructions that are no longer fully analyzable as the productive forms they began as due to their extremely high frequency. In other words, the increased use of subject/verb pairings such as *yo creo* and *yo pienso* strengthened their connection in the grammar, eventually transitioning them from individually stored lexical items to a single construction that is stored in

the lexicon.

In their investigation of lexical frequency effects in children’s variable pronoun production, Lease et al. (2022) also find that certain extremely high frequency forms act formulaic as constructions. They examine the relationship between lexical frequency and pronoun use for children in three different speech communities: (1) bilingual children in Los Angeles, CA; (2) bilingual children in Tri-Cities, WA; (3) monolingual children in Oaxaca and Querétaro, Mexico. They find that lexical frequency effects were only present for LA children, to the extent that frequent verbs favor pronoun production. They also find that highly-frequent, conventionalized forms (*yo creo* ‘I believe’ and *yo no sé* ‘I don’t know’) obscured the effects of other linguistic predictors and skewed the pronoun rates for LA children. This project highlights the differences in frequencies and their (lack of) impact across different communities of speakers, as well as the impact that extremely frequent forms have on studies of this kind. Nevertheless, this study does present findings that are in line with previous investigations of lexical frequency.

In recent literature, sociolinguists concerned with the acquisition of variable SPP use have incorporated a context-based frequency metric into their variationist studies (Brown and Shin, 2022). Brown and Shin (2022) investigate the role of contextual frequency on pronoun production by monolingual Spanish-speaking children in two age groups (ages 6-7 or 8-9). The goals of their study are to determine the extent to which conditioning contexts accumulate in the grammar for certain more frequent forms and to elucidate the acquisition of these conditioning contexts by children. The conditional frequency metric they use is defined as the rate at which each verb appears in a switch-reference or same-reference context. Results reveal similar sensitivity to conditioning context across age groups: both age groups show increased pronoun use for frequent switch-reference forms, which are posited to favor pronouns.

However, contextual frequency of switch-reference only significantly impacts pronoun production for older children and was not statistically significant for the younger children. This finding mirrors findings from studies on phonetic variation and contextual frequency: speakers (in this case, children) build up mental representations over time, storing specific details surrounding the contexts in which different lexical items appear.

Given the promising contextual frequency results in sociophonetic research, it is natural to bring contextual frequency metrics into the investigations of morphosyntactic variation. Importantly, Brown and Shin (2022) demonstrate that contextual frequency metrics fit nicely within a variationist investigation of a morphosyntactic variable. In the next section, I outline more detail surrounding UBG and provide further motivation for the contextual frequency metrics analyzed in the present dissertation.

2.3 Usage-Based Grammar

Usage-Based Grammar proposes a mental representation of language whose structure emerges through linguistic experience, i.e. perception and production. Usage-based theories assert that the mental representation of language is formed using “domain-general cognitive processes” or methods of acquisition that occur across all aspects of human mental ability, such as categorization or chunking instead of specific cognitive processes that are exclusively for language (Bybee, 2010). Crucially, usage-based theories differ from other theories of linguistic representation because they propose a dynamic linguistic system that evolves with experience. Bybee and other usage-based linguists argue against the Generative notion that language change is born exclusively out of acquisition (Chomsky and Halle, 1968), asserting that language change comes about at all stages of life through language use. Exemplar theorists propose that

all linguistic forms (e.g. lexical items, phonetic pronunciations, syntactic structures, etc.) are stored as exemplars – rich memories that contain information about both structure and meaning. Structure and meaning pairings, called constructions, are built up through categorization and exist in the grammar as syntactic *and* semantic information (Goldberg, 1995, 2006). The grammar, then, is envisioned as the aggregate of exemplar clouds.

As with any research tradition, UBG has aspects that are more robustly described and others that are actively being further developed. Usage-based models are arguably most successful in their understanding of concrete forms, with specific sensitivity to the probabilistic nature of phonological variation (Guy, 2014). However, some of the details surrounding the exemplar cloud are less fully developed. One such detail is the genesis of exemplar clouds in the first place. It is unclear how exemplar clouds are formed upon encountering a novel form. The precise characteristics of exemplars that are included in the exemplar cloud are also less clear. A third point that needs further transparency is the relationship between exemplar clouds as they exist on a spectrum from metaphors to actual neurons. These points are relevant in contextualizing the present discussion and investigation of UBG. However, the primary goals of the current study are not to provide clarity to these points, nor does this study require further consideration of these points in order to make an important contribution to our understanding of frequency in morphosyntactic variation.

Although the areas provided above are more challenging to account for, UBG does provide some details outlining their approach to mapping linguistic experience to mental representation. Each experience of a linguistic form is evaluated by the mental system based on a series of characteristics in order to determine its membership to the appropriate exemplar cloud. Then, once membership has been decided, it is evaluated once more to store that exemplar close to other exemplars of the same

category within that cloud. If, based on the relevant characteristics (e.g. F1 or roundness for a vowel), the form is near-identical to an already stored exemplar, the form acts as a reinforcement of that exemplar, boosting its frequency and detailed representation (Pierrehumbert, 2001, 2002). Over time, new exemplars are added to exemplar clouds while older exemplars are said to decay. The mental grammar is described as a web of intricate and interrelated exemplar clouds. Exemplar clouds are also highly structured based on similarities and differences not only in exemplars of the same form (such as all different phonetic productions of / ϵ /), but also in exemplar clouds of other forms. In fact, exemplar clouds are said to be organized in such a way that they form an interconnected network of exemplar clouds organized based on similarities in form and/or meaning (Pierrehumbert, 2001).

The central metaphor of the exemplar cloud guarantees that the frequency properties of linguistic forms will be important because these linguistic properties are used to organize remembered experiences of a form within the exemplar cloud. Thus, each experience reinforces existing experiences. From a usage-based perspective, high frequency features have more robust mental representations that facilitate perception, influence production, and, for certain levels of linguistic structure, even propel innovation. Usage-based theories propound that linguistic variation and change are in large part products of differences in frequency, and that these differences in frequency impact each level of linguistic structure in different ways. There is quite a bit of support for this claim: Variation and change at the phonological level is often more robust or “farther along” in higher frequency constructions than in lower frequency constructions of the same kind. Evidence of this expediting effect has been found in research on phonetic reduction phenomena across multiple languages (Hooper, 1976; Bybee, 2002; Bush, 2001; Brown and Torres Cacoullos, 2003; Díaz-Campos, 2005 and others). For instance, increased experience of a particular phrase (which corresponds

to a more robust exemplar cloud for that phrase) can often lead to greater phonetic reduction in speech production, since lexical access is facilitated by a robust representation and muscle memory allows for less articulatory effort. This is why American English speakers often produce *I am going to* as *I'm gonna*, and some speakers even reduce further, deleting the velar stop altogether (Bybee, 2010, p.39). In their study of Spanish /s/ reduction, Brown and Torres Cacoullós (2003) find that higher frequency forms are more likely to occur with /s/ reduction and even deletion. This provides support for the notion that frequency propels phonetic variation and ultimately, phonological change. Hooper (1976) finds that schwa deletion in English is more common in frequent words than in infrequent words. In her study of Spanish /d/ deletion, Bybee (2002) found that highly frequent words were more likely to present /d/ deletion intervocally than less frequent words. Highly frequent forms accumulate exemplars with phonetic reduction more quickly than low frequency forms, expediting the reduction process for those forms.

At the morphosyntactic level, however, frequency effects can manifest in different ways. In what has been referred to as a “conserving effect” (Bybee, 1985), frequency seems to reinforce existing forms in the grammar in a manner that causes them to resist reduction and even language change. For instance, certain irregular verbs can actually *evade* any sort of change due to their high frequency of use, while low-frequency irregular verbs may go through structural levelling towards regularity. One example of this is shown in the English past tense forms in the following paradigms: *sleep/slept* and *keep/kept*. Due to their high frequency of use, these past tense forms have avoided regularization unlike the less-frequent form *leap/lept*, which has regularized to *leaped* (Hooper, 1976). Usage-based studies on morphosyntactic change in Spanish have explored topics such as the ‘become’ construction (Bybee and Eddington, 2006), and the semantic shift of *andar* ‘to walk’ (Torres Cacoullós, 2000). The

extent to which this conserving effect is borne out in stable morphosyntactic variation, however, has not been studied as extensively. Nevertheless, since this conservation effect connects back to the exemplar cloud, it's possible that a similar phenomenon occurs for instances of stable variation. High frequency corresponds to increased detail in representation, which cements irregular forms in the mind. This same conserving effect, can likely be applied to stable variants, too.

UBG proposes a model of language representation that incorporates multiple components in addition to the lexical frequency discussed above. In addition to metrics on usage (i.e. frequencies of constructions with varying degrees of contextual specificity), Bybee (2010) explains that usage-based models of language also include extralinguistic details surrounding cognitive processing techniques (e.g. categorization, chunking, etc.), social context (e.g. information about the interlocutor), and language change (i.e. historical change over time) to illustrate mental representations. These features are said to be stored in the exemplar clouds, alongside linguistic forms. In other words, the contexts of use for a given linguistic form is represented in the exemplar cloud for that form. While UBG asserts that these features are present in the mental representations of these forms, the specific nature of these extralinguistic features is less clear. For instance, Pierrehumbert (2006) explains that the social factors stored in the exemplar cloud of a given form likely differ from individual to individual, since only a portion of social factors that impact production are cognitively salient in the first place, and salience of social factors varies by person. Nevertheless, these contexts of use, which include linguistic and extralinguistic factors, are said to impact speech production. The detailed nature of exemplars that usage-based frameworks outline (and which have been described here), therefore, cannot translate to one frequency metric that considers the rate of each verb within a corpus. That is to say that the verb frequency metrics explored in many previous studies on variable pronoun use

do not include as detailed information as is present in the mental representations of verb forms. These frequencies are not sensitive to the linguistic and extralinguistic constraints that are involved in the evaluation of each exemplar as it is stored in the mind.

Frequency metrics with minimal contextual specificity can be potentially misleading. For example, *estoy* ‘I am’ appears in the present corpus 1,489 times, and it disfavors pronoun production relative to the overall rate of pronoun use in the corpus with a pronoun rate of 22.83%. A monotonic reading on insights of UBG (which is more consistent with how people have previously thought about frequency as it relates to phonological variation) might suggest that the verb’s very high frequency inherently lends itself to production without a pronoun. That is to say that insofar as pronoun presence amounts to a less-reduced form, the high frequency of the verb form should promote pronoun *absence*.

- (2) (1) ahora estoy desesperado... (2) estoy en otro país sin
 (1) Now I am desperate... (2) I’m in another country without
 dinero
 money

However, when we consider some of the usage tendencies of the form, we see that the most frequent context for *estoy* (43.05%; $N = 641$) is such that there is no switch in referent *and* the preceding site of pronominal variation has a value of ‘absent’. Sentence (2) in Example (2) illustrates this context. The levels in these two pronoun-related factors are pronoun disfavoring. It would be problematic to use the verb’s frequency, in and of itself, as explanatory, since it does not capture the prevailing context-dependent usage patterns of *estoy* and could lead to a more coarse-grained interpretation of how frequency impacts the pronominal tendencies of *estoy*. Instead, context-dependent frequency as it is defined here would suggest that the low pronoun rate observed with this verb is associated with the *context* in which the verb frequently

occurs. Although contextual frequency metrics as just defined (in terms of two other factors) cannot not represent a fully elaborated exemplar-theoretic characterization of the cloud corresponding to *estoy* (which arguably includes many more details than simply switch reference/preceding pronoun frequency properties), the consideration of such a context nevertheless provides some increased information.

2.4 Better Representing Usage-Based Insights in Variationist Research

The literature described in the previous sections have provided important findings that have expanded our understanding of linguistic variation as a whole and more specifically Spanish SPPs. However, as the sections above have highlighted, multiple questions still remain. First and foremost, it is still unclear how insights from UBG might be best incorporated into variationist sociolinguistic studies of variable SPP production. While there have been some studies that have attempted to include components of UBG in variationist research, their methods for doing so have been understandably limited. For many reasons, frequency has been the obvious starting point for incorporating usage-based frameworks into variationist research. Frequency metrics integrate easily into existing quantitative techniques and are relatively simple to compute. However, additional work is needed to more faithfully represent the UBG practitioner's orientation towards frequency statistics. The research presented in this dissertation moves a step closer through an exploration of contextual frequency using a series of Contextual Frequency Metrics (Section 2.4.1).

Integrating other components of usage-based models is another puzzle that remains unsolved. For this reason, the current dissertation pursues an additional point of contact: The notion of a prefab (Section 6.4). As discussed in Chapter 1, prefabs are described as conventionalized forms, which is to say that they do not vary. This

seem to present clear problems for the sociolinguistic variable, since “two or more ways of saying the same thing” implies that the variants of a sociolinguistic variable are essentially identical in meaning but different in form. In addition to this seeming theoretical incompatibility, accounting for prefabs has its practical challenges, too. It is particularly challenging to determine what forms constitute prefabs and what forms are merely high-frequency. At what point does a form reach the level of high-frequency such that it has its own separate exemplar cloud? It is unlikely that there is a single threshold for identifying prefabs, i.e. a single frequency or rate of use at which any expression becomes prefab-like. This could present a problem for a more robust integration of UBG insights into Variationist morphosyntactic research. This topic is described in more detail in Chapter 6.

2.4.1 Contextual Frequency Metrics Proposed for the Present Study

As described in the previous section, the token frequency of verb forms provides limited detail regarding the conditioning contexts in which verbs appear. UBG asserts a mental grammar that consists of rich memories based on individual linguistic experiences. The characteristics stored in these rich memories are excluded from global frequency metrics, since these metrics are not sensitive to the linguistic and social contexts in which these verbs appear. UBG proposes that rich memories are holistic in nature, and token verb frequencies are not, which could explain the inconsistent results in previous studies on the variable. One possible method to more closely represent the details of rich memories in variationist research would be to incorporate conditional, or contextual, frequencies, which consider the rate at which a verb appears in a certain context. As described in an earlier section, this type of frequency metric has been investigated in the context of child acquisition of morphosyntactic variation, and results suggest that contextual verb frequency does impact Spanish pronoun production in children (Brown and Shin, 2022). Their findings suggest that

the effect of this context strengthens over time as children increase their linguistic experiences. As exposure to certain contexts increases, mental representations are strengthened, which in turn increases sensitivities to frequencies.

Based on the inconsistencies in previous research and the promising findings from Brown and Shin (2022), the current dissertation will investigate multiple context-based frequency metrics. The proposed context-dependent frequency metrics are based on two linguistic variables that have been shown to impact pronoun use: Switch Reference and Preceding Pronoun. The inclusion of these two linguistic predictors is a conscious one. There are two reasons why only Switch Reference and Preceding Pronoun are included as contexts for this novel frequency metric. First, they appear to be two of the most powerful predictor variables that influence overt pronoun production (Torres Cacoullos and Travis, 2018). Second, to use the nomenclature from Erker and Guy (2012), these two variables are *discursive*, instead of *systemic*, which is to say that their values change across different discourse contexts, unlike variables such as TMA or person/number. The exclusion of variables that represent social information was also intentional, since Exemplar Theory asserts that only a subset of social factors that impact production are actually cognitively salient to the level that they are stored in the exemplar cloud (Pierrehumbert, 2006).

The hypothesized pronoun-favoring frequency metric consists of the frequency at which a given verb form appears in the contexts for Switch Reference and Preceding Pronoun that favor pronoun production ('Different' and 'Present'). In contrast, the hypothesized pronoun-disfavoring frequency metric consists of the frequency at which each verb appears in the contexts for Switch Reference and Preceding Pronoun that disfavor pronoun use ('Same' and 'Absent'). The two mixed context frequency metrics consist of the frequency at which each verb appears in the contexts for Switch Reference and Preceding Pronoun that are contrasting (one pronoun favoring and

one pronoun disfavoring). Descriptions and examples of the four contexts that will be considered for the present contextual frequency metrics are described below:

- PRONOUN-FAVORING: with a switch in reference and with a value of ‘present’ at the previous site of pronominal variation (Ex. Yo hablo español. Ella habla español e inglés. ‘I speak Spanish. She speaks Spanish and English.’)
- PRONOUN-DISFAVORING: without a switch in reference and with a value of ‘absent’ at the previous site of pronominal variation (Ex. Hablo español. ∅ Quiero aprender inglés. ‘I speak Spanish. I want to learn English.’)
- MIXED CONTEXT 1: without a switch in reference and with a value of ‘present’ at the previous site of pronominal variation (Ex. Yo hablo español. ∅ Quiero aprender inglés. ‘I speak Spanish. I want to learn English.’)
- MIXED CONTEXT 2: with a switch in reference and with a value of ‘absent’ at the previous site of pronominal variation (Ex. Hablo español. Ella habla español e inglés. ‘I speak Spanish. She speaks Spanish and English.’)

While the present dissertation exclusively considers frequency metrics pertaining to Switch Reference and Preceding Pronoun, it is important to note that this configuration is by no means exhausting the logical limits of frequency metrics of this kind. That is to say, there are many combinations of other linguistic and social variables that could be considered when integrating a context-based frequency metric. Frequency metrics that consider combinations such as Semantic Category and Person/Number would be perfectly reasonable given their empirically demonstrated impact on pronoun use.

In addition to raw contextual frequency, the present dissertation utilizes an additional set of context-dependent frequency metrics titled the *Favorable Context Ratio*

(FCR) and *Disfavorable Context Ratio* (DCR). The FCR is a ratio of the frequency at which a given verb form appears in favoring contexts relative to the overall frequency of the verb. The DCR is the ratio of the frequency at which a given verb form appears in disfavoring contexts relative to the overall frequency of the verb. Unlike the other contextual frequency metrics, FCR and DCR provide insight into how often a verb's instances of use are situated in contexts that favor pronoun use and disfavor pronoun use *relative to all instances of that verb*. For example, a verb that appears in pronoun-favoring contexts 10 times and appears 20 times overall will have an FCR of 0.5. Each of these variables are described in great detail in Section 3.4.2.

The present dissertation addresses the gaps discussed above through a replication and extension of Erker and Guy (2012) with a sample size that is magnitudes larger. First, I determine whether a study with a different, larger dataset reproduces the frequency effects found in Erker and Guy (2012). I then investigate the explanatory power of Contextual Frequency Metrics (i.e. frequency metrics that are more in-line with UBG descriptions of rich memory because they are more sensitive to context) in a variable that is not phonological, hence muscle memory and reduction are not as obviously at play. Finally, I tease apart the theoretical implications of collaboration between these two frameworks.

In the following chapter, I outline the methodology for the present dissertation. In Chapter 4, I describe the nature of the data analysis and the results of the replication study. In Chapter 5, I present the results from the investigation of the six novel contextual frequency metrics. Finally, in Chapter 6, I outline the conclusions and implications of the findings of the dissertation.

Chapter 3

Methodology & Predictions

3.1 Overview

As a reminder, the research questions posed in the current thesis are as follows: (1) Does replication with a larger, different data set produce results in line with those of Erker and Guy (2012)?; (2) Does incorporating contextual frequency metrics provide clearer results?; (3) What are the theoretical implications of incorporating insights from Usage-Based Grammar into Variationist Sociolinguistic research on Spanish subject pronoun production?; and (4) To what extent can the study's results shed light on the notion of a prefab?. These research questions are motivated by the promising findings in Erker and Guy (2012) that have not yet been fully replicated.

With these research questions and motivations in mind, this chapter provides an overview of the data and methodology used for the present study. Section 3.2 describes the two corpora used for the analysis, including information on the sociolinguistic interviews and the speakers that were interviewed. Section 3.3 describes in detail the dependent variable: subject pronominal variation. Section 3.4 provides an overview of the independent variables. In Section 3.5, I outline the predictions for the independent variables. Finally, Section 3.6 details the preparation of data for the research project, and the methods of quantitative analysis.

3.2 The Corpus

As mentioned in the introductory chapter, the present study combines data from two existing corpora: The Otheguy Zentella Corpus of Spanish in New York (Otheguy and Zentella, 2012) and the Spanish in Boston Corpus (Erker, 2022). The Boston Corpus was developed with the goal of replicating the Otheguy Zentella Corpus, however, these two corpora still have some differences. Additionally, substantive data manipulation was needed in order to combine the two existing datasets and code for the new independent variables presented here. The joined and updated corpus used in the present dissertation will be referred to as the OZC-SBC Corpus.

3.2.1 Sociolinguistic Interviews

The data for this study come from sociolinguistic interviews of Spanish-speakers residing in New York City, NY and Boston, MA. The interviews were conducted in Spanish in both cities, and participants were asked a series of questions that included open-ended questions, demographic questions, and questions on language usage and attitudes. The interviews for the NYC corpus were conducted between the years of 2000 and 2005. The interviews for the Boston corpus were conducted between 2014 and 2017.

3.2.2 The Speakers

Of the 221 speakers in the present study, 80 speakers are from Boston and 141 speakers are from New York. Table 3.1 presents the speakers' Regional Origin for each corpus and overall. The percents in the table reflect the proportion of Female and Male speakers for each region relative to the total number of speakers in each corpus. The New York corpus contains two regions: Caribbean and Latin American Mainland. Speakers from Dominican Republic, Cuba, and Puerto Rico are coded as

“Caribbean” in the New York Corpus. “Mainland” speakers originate from Colombia, Ecuador, and Mexico. The Boston corpus divides “Mainland” into two more specific regional categories (“Andean” and “Central”) and also includes speakers coded as “European” and “Mixed”. The OZC-BSC therefore has five categories for Regional Origin: Andean, Caribbean, Central, European, and Mixed. Speakers are from the “Caribbean” if they are from Cuba, Dominican Republic, Puerto Rico, and coastal cities of Venezuela (namely Caracas and Maracaibo). Speakers are coded as “Andean” if they are from inland Colombia, Ecuador, Paraguay, Peru, and non-coastal parts of Venezuela. Speakers are coded as “Central” if they are from El Salvador, Guatemala, Honduras, Mexico, and Nicaragua. Speakers are coded as European if they are from Spain. Finally, speakers are coded as “Mixed” if they report that their parents are from countries in different regions (i.e. El Salvador and Dominican Republic).

The largest group of speakers overall are from the Caribbean (97 individuals), followed by Andean (65 individuals), Central (56 individuals), European (2 individuals), and Mixed (1 individual). As for the breakdown of regional origin by corpus, the Central region is the most common region for speakers in the Boston corpus (33 individuals), followed by the Caribbean (25 individuals), then Andean (19 individuals), European (2 individuals), and Mixed (1 individual). In the NYC corpus, the largest group of speakers originates in the Caribbean (72 individuals), then the Andean region (46 individuals), followed by the Central region (23 individuals). There are no European or Mixed speakers in the NYC corpus. The different representations of regional origins in the two corpora stem from the different social stratification within the cities themselves. The NYC corpus specifically focuses on the speech of informants that represent the largest Spanish-speaking populations in New York City, namely speakers from the Andean, Caribbean, and Central regions. The Boston corpus is also comprised primarily of speakers from the Andean, Caribbean, and Central

regions, but they do also include two European Spanish speakers and one speaker of mixed origin.

	Boston	NYC	OZC-BSC
	(N=80)	(N=141)	(N=221)
Regional Origin			
Andean	19 (23.8%)	46 (32.6%)	65 (29.4%)
Caribbean	25 (31.3%)	72 (51.1%)	97 (43.9%)
Central	33 (41.3%)	23 (16.3%)	56 (25.3%)
European	2 (2.5%)	0 (0%)	2 (0.9%)
Mixed	1 (1.3%)	0 (0%)	1 (0.5%)

Table 3.1: Regional origin for Speakers in the Boston Corpus, NYC Corpus, and the joint OZC-BSC Corpus.

As mentioned in Chapter 2, most of the social factors regularly examined in studies of linguistic variation (e.g. sex, age, and socioeconomic status) seem to be unreliably, or inconsistently related to variable pronoun use in Spanish. This fact actually provides further motivation for choosing pronominal variation as the source of investigation in the present study, since the best possible point of overlap between UBG and Variationist Sociolinguistics exists in a variable with influence from linguistic and cognitive factors, *not* social factors. This is largely due to the fact that UBG does not quantify social factors into their language models in the same way that variationist research does. Still, the regional information of the speakers in this corpus is presented because region of origin is associated with variable pronominal expression in Spanish (Orozco and Guy, 2008; Otheguy and Zentella, 2012 and others). Research shows that Caribbean speakers typically have higher pronoun rates than Spanish speakers from other regions. Erker and Guy (2012) do investigate the extent to which regional differences are tied to differences in frequency effects on SPP use.

They find that frequency effects are systematic across the two nationality groups in their study: Dominicans and Mexicans are sensitive to frequency in the same way.

Additional information on the speakers' age, sex, and other demographics are described in the Appendix (A.1).

3.3 The Dependent Variable

The dependent variable under investigation is PRONOUN USE. This is coded as a categorical variable with two factor levels: 'present' and 'absent'. The coding of the dependent variable was already done (Otheguy and Zentella, 2012; Erker, 2022). Numerous researchers listened to hundreds of hours of interview recordings, sectioning out all instances of finite verbs that either did occur with a subject pronoun or could have occurred with a subject pronoun but did not. Verbs were excluded as possible sites of variation if they had lexical nouns as subjects (see example (2)), if they were infinitival, or if they were gerunds¹. Tokens segmented for PRONOUN are coded as 'present' if the pronoun is produced and 'absent' if the pronoun is not produced. Consider the example transcribed in (1), which is taken from a longer utterance in Otheguy and Zentella (2012).

- (1) **Ellos llegaron** a un... a... acuerdo de que, eh, le **iban** a pagar el security...
 'They reached an... a... agreement that, uh, they were going to pay him the security...'

Both finite verbs in (1) are eligible sites of pronominal variation. The first finite verb, *ellos llegaron* 'they reached', is eligible because a subject pronoun was used, but it could have been omitted (*llegaron a un acuerdo* is perfectly grammatical). The

¹These are only some of the exclusionary factors for the verbs coded in this study. For more detailed descriptions of coding criteria for how sites of pronominal variation were identified, see Otheguy and Zentella (2012) and Erker (2022).

second finite verb, *iban* ‘they were’, is also an eligible token because a pronoun could have been used even though one was not.

- (2) Chantal quiere bailar mañana
 ‘Chantal wants to dance tomorrow’

A sentence with a lexical noun in subject position, like Example (2), would not be a possible site of SPP variation because a subject pronoun could not be used without adding focus or a pause. Based on these criteria, the dataset contains 88,001 tokens, which represent 88,001 sites of pronominal variation. We now turn to the independent variables that are analyzed in this study.

3.4 The Independent Variables

Because the current study investigates multiple kinds of independent variables, this section is divided into sections that reflect each kind of constraint. First, Section 3.4.1 outlines the linguistic variables that are investigated in the replication component of the dissertation. Section 3.4.2 then describes the LEXICAL FREQUENCY metrics under investigation, which are taken directly from Erker and Guy (2012). Finally, Section 3.4.3 defines the new CONTEXTUAL FREQUENCY metrics that will be investigated in the expansion chapter of this dissertation.

3.4.1 Linguistic Variables

As mentioned in previous chapters, one of the primary goals of this dissertation is to attempt to replicate Erker and Guy (2012). To do this, I include an investigation of all linguistic predictor variables that they considered. The following linguistic variables were taken directly from Erker and Guy (2012): MORPHOLOGICAL REGULARITY,

PERSON/NUMBER, TMA, SEMANTIC CONTENT, and SWITCH REFERENCE².

In the last decade since Erker and Guy (2012) was published, increased research has improved our understanding of each of these variables. Research on verbal morphology now suggests that morphological regularity is better understood when *not* constructed as a binary between “regular” and “irregular” (Albright et al., 2000; Albright and Hayes, 2003). Additionally, variationist research on semantic categories have found that the three verb classes may be too narrow. For these reasons, I will also investigate the efficacy of a different operationalization of the variables MORPHOLOGICAL REGULARITY and SEMANTIC CONTENT (which will be called MORPHOLOGICAL REGULARITY 2.0 and SEMANTIC CONTENT 2.0). In addition, the current dissertation will investigate the role of PRECEDING PRONOUN on pronoun use, which was not examined by Erker and Guy (2012).

Person/Number, TMA, and Switch Reference were already coded for in the NYC and Boston datasets. Morphological Regularity was added to both datasets for the current dissertation³. Semantic Content was added to the Boston dataset. Methodologies and examples for each of these linguistic variables are described below.

- **MORPHOLOGICAL REGULARITY:** Verbs were coded based on the regularity of their underlying infinitive. A verb was coded as “irregular” if its conjugated form includes an inserted velar (e.g. *yo tengo* ‘I have’ from *tener* ‘to have’), if it includes an inserted velar and change in vowel quality (e.g. *yo digo* ‘I say’ from *decir* ‘to say’), *or* if its conjugated form is seemingly “not derivable” from

²In order to remain faithful to the original Erker and Guy (2012) study, the operationalizations of certain independent variables (MORPHOLOGICAL REGULARITY and SEMANTIC CONTENT) are defined differently in this portion of the dissertation than they will be in later portions of the dissertation.

³Many of the variables included in the original Erker and Guy (2012) study require the finite verb to be included in the data frame. However, the NYC dataset did not include the finite verb. In order to code for morphological regularity, semantic content, and lexical frequency in the NYC corpus, the verbs for each token were extracted from the 141 original speaker-level excel spreadsheets. Since the original NYC study did not necessitate a specific methodology for documenting each verb, significant data tidying was needed to reduce the verb phrases to their simplest finite verb forms.

its infinitival form (e.g. *yo voy* ‘I go’ from *ir* ‘to go’). Otherwise, verbs were coded as “regular” (e.g. *hablamos* ‘we speak’ from *hablar* ‘to speak’).

- **PERSON/NUMBER:** Verbs were coded as one of six levels consisting of all possible combinations of person (i.e. first, second, and third) and number (i.e. singular and plural)⁴. For example, *yo pago* ‘I pay’ was coded as first-person singular.
- **TENSE-MOOD-ASPECT (TMA):** Verbs were coded as: indicative present (*estoy* ‘I am’), preterite (*nací* ‘I was born’), imperfect (*tenían* ‘they had’), perfect (*he visto* ‘I have seen’), future (*diré* ‘I will say’), subjunctive present (*hablen* ‘they talk’), subjunctive past (*pudiera* ‘I could’), subjunctive perfect (*hubiera sido* ‘I would have been’), periphrastic future (*van a comprar* ‘they are going to buy’), imperative (*repítes* ‘you repeat’), or conditional (*tendrían* ‘they would have’).
- **SEMANTIC CONTENT:** Verbs were coded based on their semantic class as either “mental activity” (e.g. *pensar* ‘to think’), “stative” (e.g. *estar* ‘to be’), or “external activity” (e.g. *comprar* ‘to buy’).
- **SWITCH REFERENCE:** Tokens were coded as ‘switch’ if the verbal subject of a given token is different from the previous verbal subject in the immediately preceding verb phrase. Tokens are coded as ‘no switch’ if the verbal subject has the same referent as the verbal subject in the immediately preceding verb phrase. Consider the example *Creo que ella canta* ‘I think she sings’. *Ella canta* would be coded as ‘switch’ since the previous finite verb (*creo*) has a different subject.

⁴It is important to note that second-person plural forms are extremely infrequent in the dataset. This is attributed in part to the nature of sociolinguistics interviews: Participants are almost always speaking one-on-one with an interviewer, leaving little opportunity for using second-person plural. There is also the issue that the only definitely second person plural form, *vosotros*, is regionally restricted. *Ustedes* is routinely used to refer to ‘groups of you’s’, which is arguably second person, but is treated as third person plural.

- **MORPHOLOGICAL REGULARITY 2.0:** Verbs were coded based on the regularity of its underlying infinitive. A verb was coded as “semi-irregular” if its conjugated form includes an inserted velar or if it includes an inserted velar and change in vowel quality. A verb is coded as “irregular” if its conjugated form is seemingly “not derivable” from its infinitival form. Otherwise, verbs are coded as “regular”.
- **SEMANTIC CONTENT 2.0:** Verbs were coded based on their semantic class as either “mental activity” (e.g. *pensar* ‘to think’), “estimative” (e.g. *admirar* ‘to admire’), “stative” (e.g. *estar* ‘to be’), or “external activity” (e.g. *comprar* ‘to buy’)⁵ (Enriquez, 1984; Posio, 2011; Torres Cacoullós and Travis, 2018).
- **PRECEDING PRONOUN:** A categorical variable based on the presence or absence of an overt pronoun in the preceding site of pronominal variation. A token is coded as ‘present’ if the previous clause contains a pronoun and ‘absent’ if the previous clause does not contain a pronoun.

The next section describes the three lexical frequency metrics that will be investigated in the replication portion of the current dissertation (see Chapter 4).

3.4.2 Lexical Frequencies

LEXICAL FREQUENCY is a frequency metric that considers the finite verbs that appear in the corpus. Erker and Guy (2012) operationalize lexical frequency in three different ways: RAW LEXICAL FREQUENCY, LOG LEXICAL FREQUENCY, and DISCRETE FREQUENCY. This dissertation explores the same three lexical frequency metrics. They are defined below.

⁵Although other studies have investigated more than the four semantic categories used here (Bentivoglio, 1987), this four-level operationalization was chosen due to promising results from other studies and concerns of model configuration and the proliferation of factor values.

- **RAW LEXICAL FREQUENCY:** This is a continuous variable that counts the total number of times a verb form appears in the corpus.
- **LOG LEXICAL FREQUENCY:** This is a continuous variable that consists of the natural Log-transformed Raw Frequency for each verb form.
- **DISCRETE LEXICAL FREQUENCY:** This is a categorical variable that labels verbs as “frequent” if they constitute at least 1% of the data-set ($N = 880$, in this case). Verbs are coded as “infrequent” if they occur at a rate that is less than 1%.

In the next section, I outline the novel independent variables presented in this dissertation, namely the contextual frequency metrics. Then, the predictions for these new frequency metrics are presented in Section 3.5.

3.4.3 Contextual Frequencies

The main goal of this dissertation is to analyze the role of context-dependent frequencies on pronoun use in Spanish. The hypothesis is that what some researchers have labeled cognitive variables, such as frequency metrics, are better employed if they consider certain surrounding conditions. **CONTEXTUAL FREQUENCY** in the current dissertation is defined as the rate at which each verb form appears in the following Switch Reference/Preceding Pronoun paired contexts: “Different Referent/Preceding Pronoun Absent”, ‘Different Referent/Preceding Pronoun Present’, ‘Same Referent/Preceding Pronoun Absent’, and ‘Same Referent/Preceding Pronoun Present’. This dissertation investigates six different contextual frequency metrics based on these combinations. Table 3.2 below shows the four ways that the factor values for the variables Switch Reference and Preceding Pronoun can logically combine. In the table, the target token is in the second phrase, indicated in bold font. SWITCH REFERENCE is underlined in black if the referent is the same and in orange

if the referent is different. PRECEDING PRONOUN is in green text if a pronoun is present in the previous clause and in pink text if a pronoun is absent in the previous clause.

SWITCH REFERENCE

		<u>Same</u>	<u>Different</u>
PRECEDING PRONOUN	Present	<i>Mi hermana</i> My sister <i>fue al super-</i> went to the <i>mercado. <u>Ella</u></i> supermarket. <i>compró una</i> She bought <i>galleta y <u>ella</u></i> a cookie and <i>se la comió</i> she ate it	<i>Mi hermana</i> My sister <i>fue al super-</i> went to the <i>mercado. <u>Ella</u></i> supermarket. <i>compró una</i> She bought a <i>galleta y <u>él</u> la</i> cookie and he <i>comió</i> ate it
	Absent	<i>Mi hermana</i> My sister <i>fue al super-</i> went to the <i>mercado. <u>∅</u></i> supermarket. <i>compró una</i> She bought <i>galleta y <u>ella</u></i> a cookie and <i>se la comió</i> she ate it	<i>Mi hermana</i> My sister <i>fue al super-</i> went to the <i>mercado. <u>∅</u></i> supermarket. <i>compró una</i> She bought a <i>galleta y <u>él</u></i> cookie and he <i>se la comió</i> ate it

Table 3.2: Examples illustrating the novel, context-dependent variable: CONTEXTUAL FREQUENCY.

The first four contextual frequency metrics investigated in the analysis chapter of this dissertation are based on the number of times each verb appears in these paired contexts. The ‘Different Referent/Preceding Pronoun Present’ context is arguably pronoun favoring, since it consists of both pronoun-favoring levels for the variables switch reference (‘different’) and preceding pronoun (‘pronoun present’). The ‘Same Referent/Preceding Pronoun Absent’ context is arguably pronoun disfavoring, since it contains the pronoun-disfavoring levels for Switch Reference (‘same’) and Preceding Pronoun (‘pronoun absent’). There are also two mixed contexts, which have one arguably pronoun-favoring component and one arguably pronoun-disfavoring component (‘Same Referent/Preceding Pronoun Present’ and ‘Different Referent/Preceding

Pronoun Absent’). These four frequency metrics are defined below.

- **LOG FREQUENCY-DIFFERENT-PRESENT**: This is a continuous variable that consists of the natural log-transformed frequency at which each inflected verb appears with a switch in referent (‘different’) and a value for preceding pronoun that is ‘present’.
- **LOG FREQUENCY-SAME-ABSENT**: This is a continuous variable that consists of the natural log-transformed frequency at which each inflected verb appears without a switch in referent (‘same’) and a value for preceding pronoun that is ‘absent’.
- **LOG FREQUENCY-SAME-PRESENT (Mixed Context 1)**: This is a continuous variable that consists of the natural log-transformed frequency at which each inflected verb appears without a switch in referent (‘same’) and a value for preceding pronoun that is ‘present’.
- **LOG FREQUENCY-DIFFERENT-ABSENT (Mixed Context 2)**: This is a continuous variable that consists of the natural log-transformed frequency at which each inflected verb appears with a switch in referent (‘different’) and a value for preceding pronoun that is ‘absent’.

Because the present dataset is quite large ($N_{tokens} = 88,001$), there are extreme differences in verb frequency across inflected forms. For example, the most frequent verb form (*sè* ‘I know’) appears 3,062 times. In contrast, there are 4,013 verbs that appear only one time in the corpus. These different frequencies of occurrence could pose a non-trivial problem for investigating contextual frequency effects, since verbs with overall higher frequencies will undoubtedly have higher raw values for the contextual frequencies. For this reason, the present dissertation also includes two

proportion-based variables: FAVORING CONTEXT RATIO (FCR) and DISFAVORING CONTEXT RATIO (DCR). These are defined below:

- FAVORING CONTEXT RATIO: The proportion at which each verb form occurs in the ‘Different Referent/Preceding Pronoun Present’ context (arguably a favoring context) relative to the total number of its occurrences.
- DISFAVORING CONTEXT RATIO (DCR): The proportion at which each verb form occurs in the ‘Same Referent/Preceding Pronoun Absent’ context (arguably a disfavoring context) relative to the total number of its occurrences.

Unlike the four CONTEXTUAL FREQUENCY metrics, which report log frequencies, this variable adjusts for differences in overall rates in order to facilitate comparisons between forms with different overall frequencies. For example, of the 3,062 instances of *sé* ‘I know.indicative’, 556 occur in the context described here as ‘Different Referent/Preceding Pronoun Present’. Therefore, the FCR for *sé* is 0.18. *Hacen* ‘they do.indicative’ appears in the ‘Different Referent/Preceding Pronoun Present’ context 41 times and appears in the corpus 223 times overall. Regardless of the differences in overall frequency, the FCR for *hacen* is the same as the FCR for *sé*: 0.18. One question that does arise is whether the potential effect of a (dis)favoring context ratio is itself contingent on overall frequency. To situate this in the example just presented, exemplar theory would argue that although *sé* and *hacen* have the same FCR, they would not be shaped by the contextual occurrence properties in the same way due to their differences in overall frequency. If we speculate on the respective exemplar clouds for these forms, presumably the cloud for *sé* is much larger, and contains much more detail than the cloud for *hacen*. Given these differences, the potential impact of the FCR could possibly be more robust for forms with larger exemplar clouds like that corresponding to *sé*. This question will be explored in greater detail in Chapter 4.

Like the FCR, the DCR adjusts for differences in overall rates in order to account for verb forms that appear in low frequencies. For example, *sé*, appears in the ‘Same Referent/Preceding Pronoun Absent’ context 818 times, and it appears in the entire corpus 3,062 times, which means the DCR for *sé* is 0.27. The verb form *conozca* ‘I know.subjunctive’ appears in the ‘Same Referent/Preceding Pronoun Absent’ context 3 times and it appears only 11 times overall. Nevertheless, the DCR for *conozca* is also 0.27. Importantly, the FCR and the DCR are *not* simply mirror images of each other. They are individual proportions that consider Frequency-Different-Present or Frequency-Same-Absent contexts relative to overall verb rate for each form, and they exclude mixed context frequencies. In other words, the FCR takes into account only one of the four logical combinations of Switch Reference/Preceding Pronoun (Frequency-Different-Present) and ignores the other three contexts (Frequency-Same-Absent, Frequency-Same-Present, and Frequency-Different-Absent). In contrast, the DCR considers the Frequency-Same-Absent context, while excluding the other three contexts (Frequency-Different-Present, Frequency-Same-Present, and Frequency-Different-Absent).

The inclusion of the contextual frequency metrics described above provides a more faithful representation of Usage-Based Grammar insights within a variationist framework. I investigate the significance, if any, of each of these variables individually and combined with the set of the linguistic variables. These metrics aim to shed light on the extent to which rich memory (operationalized here as context-dependent frequencies) interact with linguistic variables to influence Spanish variable pronoun production.

3.4.4 Summary of the Independent Variables

Table 3.3 provides a summary of the independent variables analyzed at each part or phase of the study. The table includes the variable name, the kind of variable

it is, and the levels for that variable. The table is divided into three parts: the first part lists the linguistic variables (which include the linguistic constraints from Erker and Guy (2012) and three new constraints), the second part includes the lexical frequency metrics taken from Erker and Guy (2012), and the third part reflects the novel contextual frequencies presented in this dissertation.

The next section outlines the predictions for how the novel contextual frequencies will interact with pronoun use. The predictions for the contextual frequency metrics stem from previous variationist research that has investigated contextual frequency and from UBG theories of frequency effects.

3.5 Predictions for Contextual Frequencies

From a usage-based perspective, it would be predicted that the overall pronoun rate of a verb form should be largely due to the pronoun-favoring-or-disfavoring properties of the context in which it occurs most frequently. That is to say that the most-frequent context will dominate the overall pronominal tendencies of the verb in *all* instances of use, assuming that Preceding Pronoun and Switch Reference are indeed the strongest predictor variables. Another thing to consider is whether the potential effect of contextual frequencies are dependent on overall frequency of occurrence. Since UBG asserts that more frequent forms have more robust mental representations, it could be the case that only verb forms with high enough frequencies will accumulate a sufficiently robust mental representation that is sensitive to the impact of contextual frequencies. What constitutes a “high-enough frequency” has yet to be determined. Nevertheless, it would be predicted that verbs that appear more frequently in ‘Different Referent/Preceding Pronoun Present’ contexts would generally be produced with higher overt pronouns across all contexts, since ‘Different Referent/Preceding Pronoun Present’ contexts are presumably SPP favoring with respect to the value of the

	Type	Levels
Linguistic Variables		
MORPHOLOGICAL REGULARITY	Categorical	regular, irregular
PERSON/NUMBER	Categorical	first singular, second singular, third singular, first plural, second plural, third plural
TMA	Categorical	indicative present, preterite, imperfect, perfect, future, subjunctive present, subjunctive past, subjunctive perfect, periphrastic future, imperative, conditional
SEMANTIC CONTENT	Categorical	mental activity, stative, external activity
SWITCH REFERENCE	Categorical	same referent, different referent
MORPHOLOGICAL REGULARITY 2.0	Categorical	regular, semi-irregular, irregular
SEMANTIC CONTENT 2.0	Categorical	mental activity, stative, estimative, external activity
PRECEDING PRONOUN	Categorical	preceding pronoun present, preceding pronoun absent
Lexical Frequency Metrics		
DISCRETE FREQUENCY	Categorical	frequent, infrequent
RAW LEXICAL FREQUENCY	Continuous	$N \geq 1$
LOG LEXICAL FREQUENCY	Continuous	$\log(N) \geq 0$
Contextual Frequency Metrics		
CONTEXTUAL FREQUENCY	Continuous	$\log(N) \geq 0$
FAVORING CONTEXT RATIO	Continuous	$0.00 \leq N \leq 1.00$
DISFAVORING CONTEXT RATIO	Continuous	$0.00 \leq N \leq 1.00$

Table 3.3: Summary of all Independent Variables

two properties upon which this variable is based. A UBG approach would also predict that verbs that appear most frequently in ‘Same Referent/Preceding Pronoun Absent’ contexts would generally be produced with lower rates of overt pronouns. This context is presumably pronoun disfavoring in terms of the value of the two properties that comprise this variable.

As for the two mixed contexts (“Different Referent/Preceding Pronoun Absent’ and ‘Same Referent/Preceding Pronoun Present’), the predictions are much less straightforward, since these contexts each contain one value that is arguably pronoun favoring and one value that is arguably pronoun disfavoring. If these two properties hold even weight, which is to say they contribute equally in pronoun presence/absence, we might predict that verbs that appear most frequently in these contexts would generally be produced with overt pronouns at a rate that is neither high nor low, but in the middle. However, if either of these values outweighs its pair in respect to influencing overt pronoun production, we might predict that verbs that appear in this context will be produced with overt pronouns at rates that mirror the direction of that value. For example, if ‘different’ outweighs ‘absent’, we would predict that verbs that appear most frequently in the “Different Referent/Preceding Pronoun Absent’ mixed context will correspond to higher overt pronoun rates, since a ‘different’ referent corresponds with overt pronoun production. In contrast, if ‘present’ outweighs ‘same’, a verb that appears most frequently in ‘Same Referent/Preceding Pronoun Present’ contexts will correspond to higher overt pronoun rates, since pronoun presence in the previous site of pronominal variation supports overt pronoun production. Since the CONTEXTUAL FREQUENCY MEASURES consist of four individual frequencies, it is unclear whether each of these will significantly impact pronoun production. Nevertheless, the current study will investigate the constraining power of all four.

Additionally, I predict that speakers will favor overt pronouns for verbs with higher

FCRs. Since the FCR is a ratio of the frequency at which a verb occurs in the ‘Different Referent/Preceding Pronoun Present’ context over the overall frequency of the verb, higher FCR corresponds to a higher occurrence in ‘Different Referent/Preceding Pronoun Present’ contexts, which should suggest higher pronoun rates. In contrast, speakers are predicted to produce lower rates of overt pronouns for verbs that have higher DCRs. This is because the DCR is a proportion that considers the rate at which each verb appears in pronoun-disfavoring contexts compared to their overall rates in the corpus. A verb with a higher DCR therefore appears more often in ‘Same Referent/Preceding Pronoun Absent’ contexts. Ultimately, I predict that these enriched frequency metrics will provide increased clarity into the effects of frequency on variable pronoun production in Spanish.

3.6 Statistical Methods

The analysis of the present dissertation occurs in two parts, each with their own statistical analyses. The first part, reported in Chapter 4, consists of a replication of Erker and Guy (2012). This replication diverges from the original in its use of multivariate analyses instead of the univariate analyses Erker and Guy (2012) carry out. The second phase of the analysis expands upon the replication by investigating the extent to which contextual frequencies impact pronoun use in the OZC-BSC Corpus (see Chapter 5).

3.6.1 Replicating Erker and Guy (2012)

As mentioned, the Erker and Guy (2012) study utilized univariate analyses (i.e. *t*-tests, correlations, and ANOVAs). The statistical methods I use in this portion of the study are different from the original. To determine the main effects of the continuous variables (*RAW FREQUENCY* and *LOG FREQUENCY*), Pearson’s correlations were run on pronoun rates and each continuous variable. To determine the main effects of

the independent variables (DISCRETE FREQUENCY, MORPHOLOGICAL REGULARITY, and SWITCH REFERENCE, PERSON/NUMBER, TMA, and SEMANTIC CONTENT), logistic mixed effects regression models were run using the lme4 package in R (Bates et al., 2015). Logistic mixed effects models were also run to investigate the extent to which there were any interactions between discrete frequency and the other linguistic predictors. At the end of the replication chapter, I investigate the potential main effects of the three additional linguistic predictors (MORPHOLOGICAL REGULARITY 2.0, SEMANTIC CONTENT 2.0, and PRECEDING PRONOUN) and their potential interactions with discrete frequency. More detail surrounding the analysis for the replication is presented in Chapter 4.

Potential limitations

Because of the sheer size of the dataset in this dissertation project, and because of the nature of the dependent variable, the statistical analyses employed in the original study were not best suited for this replication. The statistical methods carried out in Erker and Guy (2012) were the state of the art at the time the original study was published. It is possible that different statistical methods provide different results. However, any replication can produce different results even if the participants, variables, and statistical methods are identical. The broader goal for this portion of the study is to investigate whether lexical frequency acts as an amplifier or activator as was shown in Erker and Guy (2012), and this goal can be accomplished despite the differences in statistical methods.

3.6.2 Extending Erker and Guy (2012)

In the second phase of the study, I leverage multivariate analyses for the extension of Erker and Guy (2012). The expansion contains token-level analyses as well as rates-level analyses. I utilize logistic mixed effects regression models that include

verb as a random effect, one predictor as a fixed effect, and pronoun production as the response variable. For the three linguistic predictors, I also look at the extent to which an interaction is occurring with the predictor and discrete frequency (as was done in the replication). For the contextual frequency metrics, I carry out a correlation matrix to assess the relationship between overall frequency and the four contextual frequency metrics. Then, I investigate the Pearson's Correlations between each metric and their corresponding rates of pronoun use. For FCR and DCR, I also investigate the potential for an interaction between each ratio-based metric and Log Frequency using logistic mixed effects regression models. More detail surrounding the analysis for the expansion is presented in Chapter 5.

Chapter 4

Replicating Erker & Guy (2012): Return to Lexical Frequency

This chapter presents the replication portion of the dissertation. This replication includes multiple lexical frequency metrics along with five linguistic predictors. Although the primary intention is to replicate Erker and Guy (2012) as closely as possible, the results do not consist of exactly the same quantitative analysis, as mentioned in the previous chapter. Instead of t-tests, chi-squared tests, and ANOVAs, the analysis below uses mixed effects logistic regression models to statistically evaluate the main effects and interactions of the linguistic predictors. Due to the potential for within-verb and between-verb differences, mixed effects models are necessary. The findings overall are in line with those found in Erker and Guy (2012): Discrete Frequency acts as either an amplifier or an activator for other linguistic predictors. However, the results deviate from Erker and Guy (2012) as well. The primary difference is that there is stronger (though still inconclusive) evidence that higher frequency is directly associated with more pronoun use, which is to say there is a strong positive correlation. Nevertheless, further investigation reveals that high frequency, in and of itself, is a poor predictor of pronoun rates, especially among the highest frequency forms.

In Section 4.9 I outline the results of the updated operationalization for the variables morphological regularity and semantic category, as well as the addition of the variable Preceding Pronoun as a linguistic predictor. Results indicate that Preceding

Pronoun significantly impacts pronoun production and the updated operationalizations of morphological regularity and semantic category better account for pronominal variation.

4.1 General Overview of the Data

As mentioned in earlier chapters, the goal of the replication is to answer the following research question: Does replication with a larger, different data set produce results in line with those of Erker and Guy (2012)? To answer this question, this section will focus on investigating the five linguistic constraints and three lexical frequency measures from their study. Each frequency metric was defined based on the occurrence of the verb forms within the corpus ($N_{(Erker\&Guy,2012)} = 4,916$). When setting out to analyze raw frequencies, Erker and Guy (2012) note some of the methodological challenges that arise. One such challenge is the nature of frequency distributions for words, which are known to follow Zipf’s law (Zipf, 1945). Zipf’s law highlights the non-normal distribution of word frequency as it relates to word rank. This is visualized in Figure 4-1, which plots the raw counts for all verb forms¹ in the current corpus from most to least frequent. As the figure shows, there are very few high frequency forms (at the left of the figure), but very many low frequency forms (trailing off along the x-axis).

As previously discussed (Chapter 3.4.2), frequency in this replication is measured in the same manner as Erker and Guy (2012). This means that RAW FREQUENCY, which consists of a simple count of each occurrence of a given verb form in the corpus, is included in the investigation. To account for the extreme differences in raw frequencies, we employ a log frequency metric like Erker and Guy (2012): the natural log transformation of the raw frequency for each verb. Using a log transformation on

¹For visualization purposes, verbs that occurred less than 3 times are excluded from the figure.

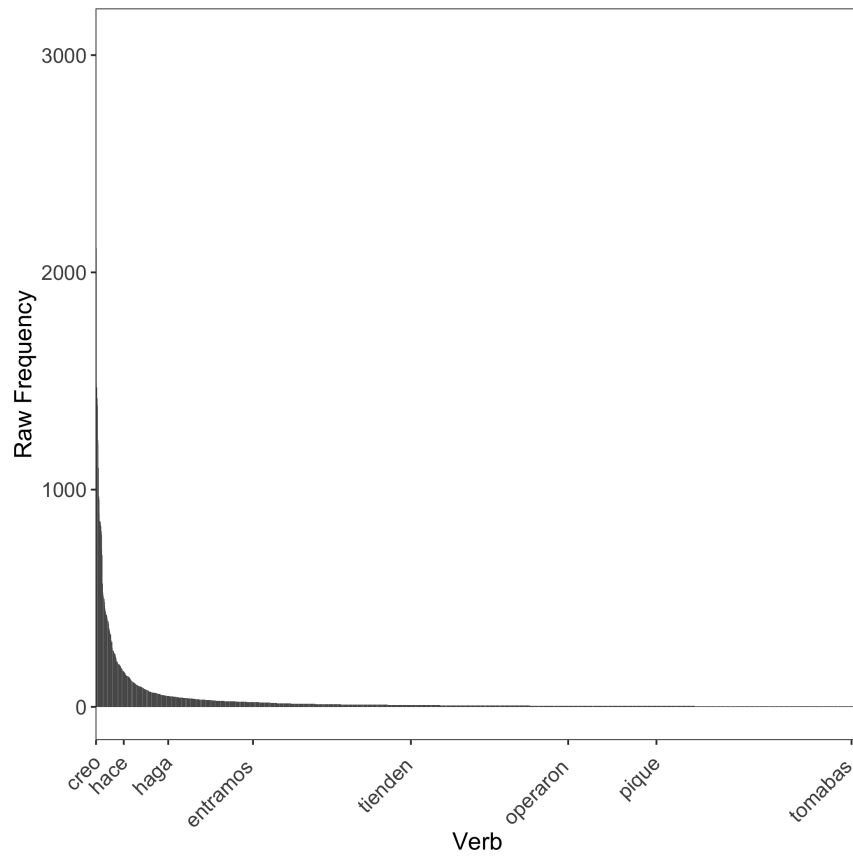


Figure 4-1: Zipf's law illustrated through raw frequency distribution of verb forms in the OZC-BSC Corpus.

continuous data allows for the data to more closely resemble a normal distribution, which can increase the validity and interpretability of statistical analysis. Finally, DISCRETE FREQUENCY is the categorical frequency metric investigated in Erker and Guy (2012). This metric assigns a label of “frequent” to all verbs that constitute at least 1% in the corpus. All other verbs are assigned a label of “infrequent”. This means that verb forms need to appear at least 880 times in the current corpus to be considered a frequent form. Frequent forms in Erker and Guy (2012) appeared at least 49 times.

Table 4.1 shows the verb forms that comprise at least 1% of the corpus for the present dissertation² and for Erker and Guy (2012). These forms were therefore labelled as “frequent”. For the OZC-BSC Corpus, their combined frequency of occurrence is 19,290, and they make up 21.9% of the corpus. The remaining 78.1% of the tokens in the corpus were coded as “infrequent”.

Generally, there is quite a bit of overlap in the frequent forms in the current dissertation and the frequent forms in Erker and Guy (2012). Only two frequent forms on the OZC-BSC side are not present in Erker and Guy (2012) (*tiene* ‘he/she have’ and *son* ‘they are’) . Conversely, *ves* ‘you see’, *era* ‘I/he/she/you was/were’, and *fui* ‘I was/I went’ accounted for at least 1% of the corpus in Erker and Guy (2012), but did not in this dissertation.

A t-test was carried out to compare the mean frequencies for frequent and infrequent verbs in the dissertation data. Results indicate that these two groups have significantly different average raw frequencies, suggesting that dividing the corpus at the 1% threshold creates two meaningfully different groups in terms of raw frequency. For infrequent verb forms, the mean raw frequency is 231.6, and the standard deviation is 287.4. The mean raw frequency for infrequent forms is somewhat misleading,

²Forms marked with an asterisk represent surface forms that are identical across different person/number combinations. For example, *tenía* can be first-person singular or third-person singular.

OZC-BSC				Erker & Guy (2012)			
FORM		COUNT	%	FORM		COUNT	%
sé	‘I know’	3,062	3.5	creo	‘I believe’	204	4.1
creo	‘I believe’	2,952	3.4	sé	‘I know’	148	3.0
tengo	‘I have’	2,111	2.4	digo	‘I say’	117	2.4
estoy	‘I am’	1,489	1.7	tengo	‘I have’	92	1.9
digo	‘I say’	1,471	1.7	sabes	‘you know’	82	1.7
estaba*	‘I/he/she/you was/were’	1,421	1.6	ves	‘you see’	68	1.4
tenía*	‘I/he/she/you had’	1,387	1.6	estaba*	‘I/he/she/you was/were’	67	1.4
sabes	‘you know’	1,228	1.4	estoy	‘I am’	61	1.2
es	‘he/she/you are’	1,100	1.2	tenía*	‘I/he/she/you had’	61	1.2
soy	‘I am’	968	1.1	era*	‘I/he/she/you was/were’	59	1.2
tiene	‘he/she have’	955	1.1	soy	‘I am’	58	1.2
son	‘they are’	891	1.0	fui*	‘I was/I went’	54	1.1
				es	‘he/she/you are’	49	1.0
TOTALS		19,032	21.6	TOTALS		1,120	22.8

Table 4.1: Frequent verb forms

since there is quite a bit of variation in the frequencies within that group. For frequent forms, the mean raw frequency is 1,903, and the standard deviation is 812.6. There is also quite a high range of variation within the group labelled “frequent”. Overall, these results, presented in Table 4.2, show a smaller disparity of frequencies between frequent and infrequent forms than was found in Erker and Guy (2012). The average raw frequency for infrequent forms in their study is 8.2, and their average raw frequency for frequent forms is 108.1 ($t(4,914) = 106, p < 0.001$).

The mean raw frequency for infrequent forms in the present study may seem surprisingly high considering that it is nearly double the average frequency for *frequent* forms in Erker and Guy (2012). However, given the size of the corpus, a mean raw frequency of 231.6 signifies that each low frequency verb form appeared approximately once per interview on average ($N_{speakers} = 221$). While alternative cut-off points for the frequent/infrequent divide are likely worth exploring, that is not within the scope

DISCRETE FREQUENCY	MEAN RAW FREQUENCY	<i>N</i> TOKENS
Infrequent (verbs that are each < 1% of data)	231.6	68,969
Frequent (verbs that are each ≥ 1% of data)	1,903.4	19,032
t(20,350) = 279.05, <i>p</i> < 0.001		

Table 4.2: Mean raw frequency: frequent vs. infrequent forms

of this stage of this analysis. In order to maintain as close a replication as possible, this chapter retains the methods of Erker and Guy (2012).

4.2 Main Effects of Lexical Frequency

In order to determine the extent to which lexical frequency impacts pronoun production, we must first consider, as Erker and Guy (2012) do, the extent to which there is evidence of a direct statistical relationship between pronoun use and raw, log, and Discrete Frequency. Figure 4-2 presents the percent of pronoun production observed for each raw frequency. There are very many low frequency forms and very few high frequency forms, such that the points on the lower end of the x-axis correspond to many different verb forms and the points on the upper end of the Raw Frequency range correspond to unique verb forms. For instance, while there are 183 different verb forms with a frequency value of ‘6’ (e.g. *canto* ‘I sing’, *crecimos* ‘we grew up’, *dependo* ‘I depend’, *invitan* ‘they invite’, etc.), this value is represented by a single point on the plot. In contrast, there is only one form with a frequency value of 1,100 (*es* ‘he/she/you are’). This figure is consistent with an analogous visualization by Erker and Guy (2012): the vast majority of points are located on the left side of the plot. The relationship between raw frequency and pronoun rate is obscured due to extreme differences in verb frequencies such that there are many less-frequent

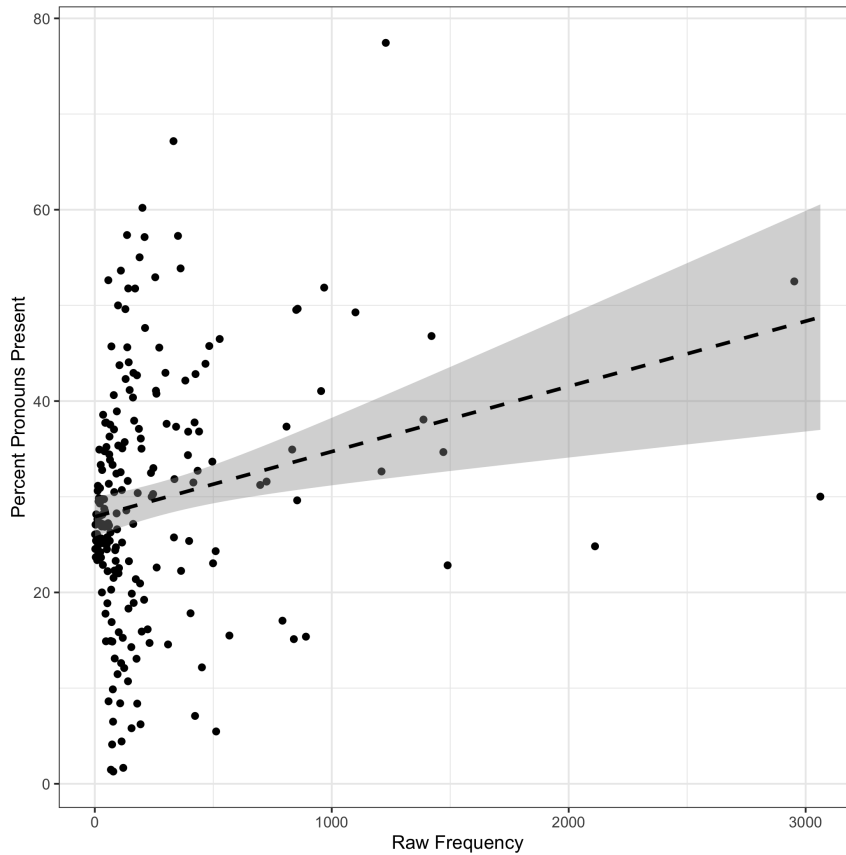


Figure 4.2: Percent pronouns present within each raw frequency

forms and very few highly-frequent forms. Nevertheless, a correlation statistic that considers pronoun rate and raw overall frequency indicates a moderate correlation: $r(87,999) = 0.31, p < 0.001$. This suggests that higher raw frequency significantly increases the likelihood of pronoun production. This finding is quite a bit stronger than that of Erker and Guy (2012), who found a significant effect, but a very weak correlation ($r(4,916) = 0.057, p < 0.001$).

Since frequency data of this kind exist in a power law distribution (i.e. very few high-frequency forms and very many low-frequency forms), visualizations and correlation tests run on raw frequency provide limited insight. Log frequency has the potential to present clearer results since the log can account for extreme differences in frequencies across verbs. Figure 4.3 plots the pronoun rates by overall log frequency.

Like the raw frequency plot above (Figure 4.2), by creating bins of verb forms at each log frequency value, some points included multiple *different* forms. Here, we can better see the relationship between frequency and pronoun production than was observable in the raw frequency plot.

The linear fit of Figure 4.3 suggests a consistent upward trend of pronoun rates as log frequency increases. It is clear, though, that many high-frequency forms are poorly predicted by the linear fit in the figure. Although there is an observable increase in pronoun rates, higher log frequency seems to actually correspond to increased variation in pronoun rates. As the figure shows, pronoun rates seem to steadily rise as the log frequency increases until it reaches a log frequency of around 4. After this threshold is reached, the level of dispersion in the pronoun rates for verb forms increases substantially, and the majority of verb forms fall outside of the confidence interval (shaded in grey). Erker and Guy (2012) provide a possible explanation for this: higher frequency verbs possess individual pronominal tendencies that are unique to each verb, while lower frequency verb forms may operate at a baseline pronoun rate.

A Pearson's correlation statistic returns a significant correlation between log frequency and pronoun rate that is stronger than raw frequency ($r(87,999) = 0.34, p < 0.001$). Interestingly, the log frequency findings presented here are inconsistent with those found in Erker and Guy (2012). Statistical analysis revealed that the log frequency measure in their dataset was related to pronoun production in an inverse way: higher log frequency correlating with lower pronoun rates. This finding was very weak and ultimately not seriously considered in light of their other findings. The correlation results in the present study continue to support the notion that high lexical frequency impacts pronoun production in a positive direction. Higher frequency (when assessed via a log transformed continuous measure) corresponds to more overt pronoun use

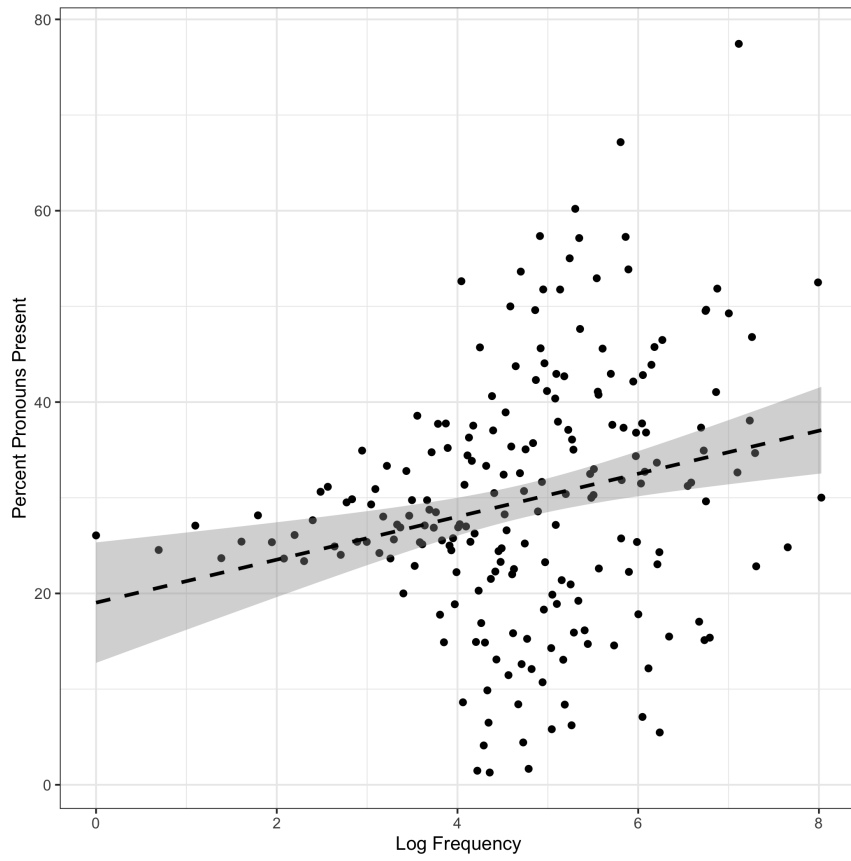


Figure 4.3: Percent pronouns present within each log frequency

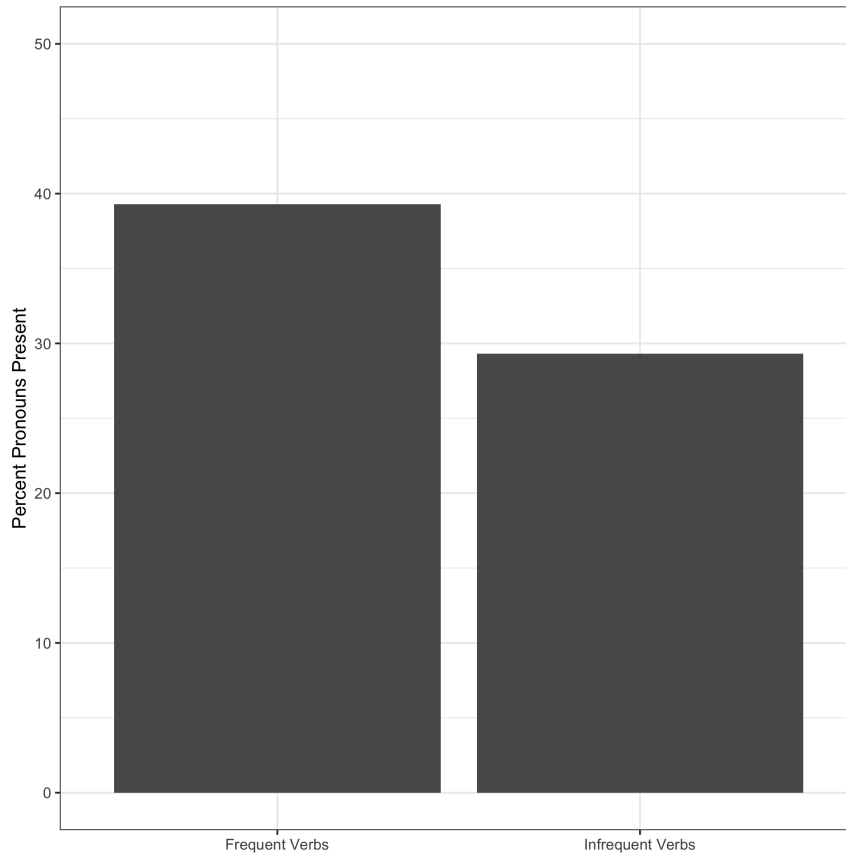


Figure 4-4: Frequent and Infrequent forms by percent SPPs present

when calculated as the overall rate of pronoun use for all the verbs with a given log frequency value. However, when we observe the data plotted in Figure 4-3, it is clear that the correlation results do not capture everything that is happening in the data. Since the majority of forms fall outside of the linear fit in Figure 4-3, it is evident that log frequency is not consistently impacting pronoun use in the direction that is suggested by the correlation statistic. We now turn to the third and final frequency metric analyzed in Erker and Guy (2012): Discrete Frequency.

Figure 4-4 plots the pronoun rates for categorically frequent forms (see Table 4.1) and categorically infrequent forms. Frequent verbs have an average pronoun rate of 39.3%, while infrequent verbs have an average pronoun rate of 29.3%. This finding is quite different from the result demonstrated in Erker and Guy (2012): they find

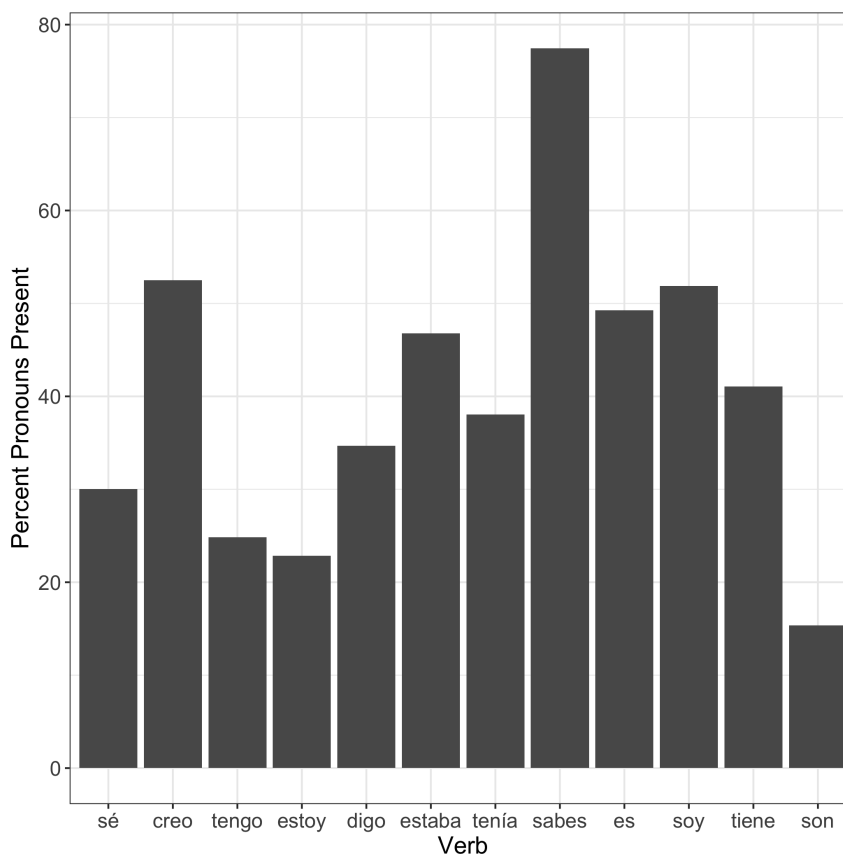


Figure 4·5: Percent SPPs present for highest-frequency forms

only a 1% difference in pronoun rates for frequent verb forms (32%) and infrequent verb forms (31%). A chi-square confirms the significant difference in pronoun rates between these two groups: $\chi^2(405, N = 88,001) = 88,001, p < 0.001$. Erker and Guy (2012) also saw a significant result from their t-test comparing pronoun rates for frequent and infrequent forms ($t = 5.6, p < 0.001$). This finding coupled with the correlation statistics for raw and log frequency seem to suggest a positive correlation between pronoun rate and overall frequency. However, when we look at the data visually (in Figures 4·2 and 4·3), we see that a positive main effect of frequency is not really impacting pronoun use linearly. This conclusion becomes even clearer when we look at overt pronoun rates for the high-frequency forms. Consider Figure 4·5.

The main effects found for all three Erker & Guy lexical frequencies could lead to

the expectation that all twelve high-frequency forms will have relatively high overt pronoun rates. Additionally, the statistically significant positive correlation between frequency and pronoun production would predict a steady tapering of pronoun rate as frequency decreases. However, Figure 4-5, which plots the percent of pronouns present for each of the high-frequency forms from highest to lowest frequency, suggests otherwise. This figure shows quite a bit of variation between individual highly frequent forms, which contradicts the expected increase in pronoun rates that is suggested by statistical analyses. For example, *sabes* ‘you know’ has the highest pronoun rate (77.4%) but is only the sixth most frequent form ($N_{sabes} = 1,228$). This visual coupled with the previous findings for raw and log frequency reaffirm that the pattern in the data is nonlinear. Further, a one-way ANOVA shows significant differences in pronoun rates for these forms ($F = 6.818e + 27, p < 0.001$). This finding again falls in line with that of Erker and Guy (2012), who also found significant differences in SPP production for their high-frequency forms.

Taken together, these results provide some differences from Erker and Guy (2012), but there is no more clarity on the role of lexical frequency than they reported. Statistical analyses reveal significant main effects for raw frequency, log frequency, *and* Discrete Frequency on pronoun use. However, the notable differences in pronoun rates for individual high-frequency forms are in clear conflict with these effects. If verb frequency affects pronoun production like other predictors (i.e. presumably impacting verb forms uniformly), we would expect a steady increase in pronoun use as frequency increases. However, this is not what Figures 4-2, 4-3, and 4-5 depict. Instead, these figures reveal that individual, high-frequency verb forms seem to have their own unique pronominal tendencies. This trend is reinforced when we expand the description of “frequent” to include any forms that comprise at least 0.5% of the corpus data ($N = 440$, in this case)³. Figure 4-6, which presents the pronoun rates for

³As previously mentioned, investigating the efficacy of other Frequent thresholds is outside the

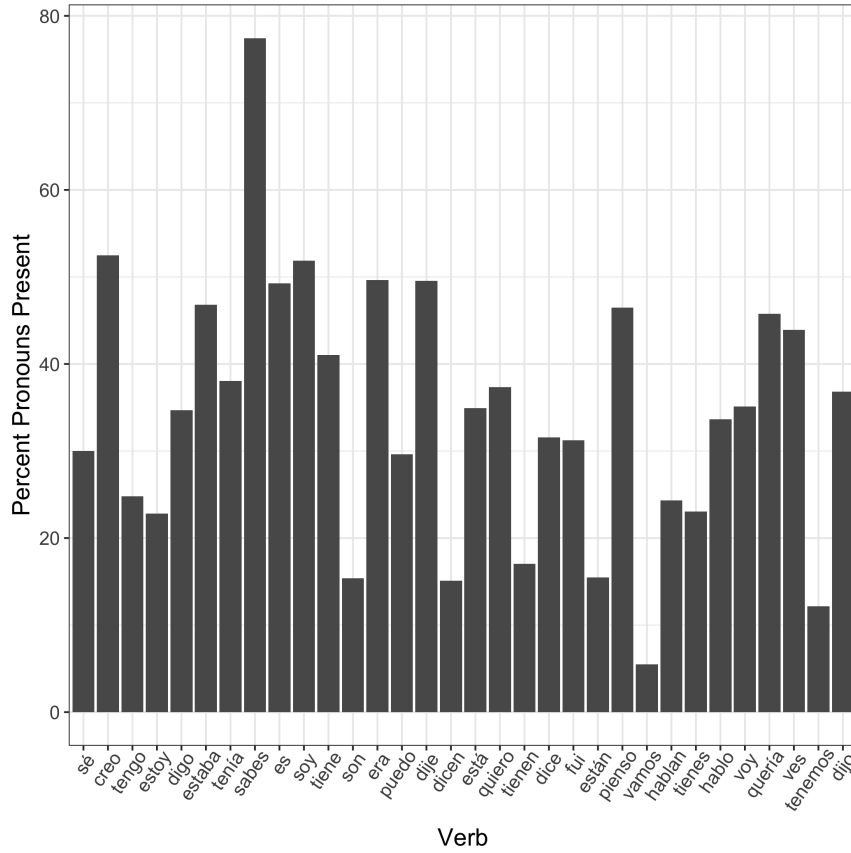


Figure 4.6: Percent SPPs present for verb forms that appear at least 440 times in the corpus (comprising 0.5% of the corpus)

the 32 most-frequent verb forms, reaffirms that highly-frequent forms do not present obvious linear trends in pronoun rates. Further exploration of these findings will be presented in the Discussion section of Chapter 5.

We now turn to the additional linguistic predictors that were investigated in Erker and Guy (2012). These can provide additional insight into the frequency effects that were reported above. The analyses of these linguistic conditioning factors are divided into two parts: (1) ‘main effects’, which assess direct relationships between the dependent variable (pronoun use) and the independent variable, and (2) interactions, which

scope of this dissertation. This figure is included purely for expository purposes and no statistical analysis is run on alternative Frequent thresholds.

explore the relationship between the independent variable and Discrete Frequency, if any, and its effect on the dependent variable. Investigating the additional linguistic predictors in this way allows for the current dissertation to examine whether Discrete Frequency acts as an amplifier or activator of the effects of other linguistic predictors, which was found to be the case in Erker and Guy (2012).

4.3 Morphological Regularity

As discussed in Chapter 3.4, Morphological Regularity à la Erker and Guy (2012) consists of a binary variable, where verbs are categorized as either “regular” or “irregular”. First, let us investigate the extent to which Morphological Regularity impacts pronoun production directly.

	<i>N</i> VERBS	% OVERT PRONOUNS
Regular Forms	45,675	32.2
Irregular Forms	42,326	31

Table 4.3: Morphological regularity of the verb

Morphologically regular verb forms are more likely to occur with overt pronouns than irregular verb forms (see Table 4.3). However, there is minimal difference in pronoun rates between regular (32%) and irregular verb forms (31%). This finding is in line with Erker and Guy (2012), although their data presented even stronger differences: morphologically regular verb forms had 7% higher overt pronoun rates (37%) than irregular forms (30%).

A mixed effects logistic regression model was run with MORPHOLOGICAL REGULARITY as a fixed effect and VERB as a random effect to evaluate the effect of the linguistic predictor (pronoun presence/absence⁴) on its own. Since the analysis in the

⁴“Pronoun” is used in model configurations as a placeholder for the binary dependent variable pronoun presence or absence.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.14	0.04	-27.67	< 0.001***
Regular	-0.07	0.05	-1.59	0.11

Table 4.4: Main effect of Morphological Regularity; Model configuration: `glmer(Pronoun ~ Morphological Regularity + (1 | Verb))`; Reference level: Irregular

current study does not consider any social factors, `SPEAKER` was not included as a random effect in this model or any subsequent mixed effects models. Presented in Table 4.4, model results indicate that there is not a significant main effect of morphological regularity on pronoun production despite the (small) difference in overall rates for regular vs. irregular forms ($\beta = -0.074, p = 0.112$)⁵. Further, the Marginal R^2 , which corresponds to the amount of observed variation in the data that is accounted for by the fixed effects in the model, is 0.000 for this model, indicating no effect of morphological regularity. In contrast, the Conditional R^2 , which corresponds to the amount of variation in the data that is accounted for by random effects, is 0.195. These results are inconsistent with that of Erker and Guy (2012), since their t-test reveals a significant difference in pronoun rates for regular and irregular forms.

4.3.1 Interaction with Discrete Frequency

Figure 4.7 presents pronoun rates when considering morphological regularity *and* Discrete Frequency. In the previous section, we report no significant main effect of morphological regularity on pronoun production. However, the figure suggests there may be more to this story. There are clear differences in pronoun rates for irregular and regular forms when accounting for Discrete Frequency. Frequent regular forms show a large increase in pronoun production compared to infrequent regular forms

⁵A second mixed effects model was run with `VERB` and `SPEAKER` as random effects to ensure maximum reduction in type I error. The results from that model were not qualitatively different from the results in the model presented in Table 4.4, so they were excluded from the analysis.

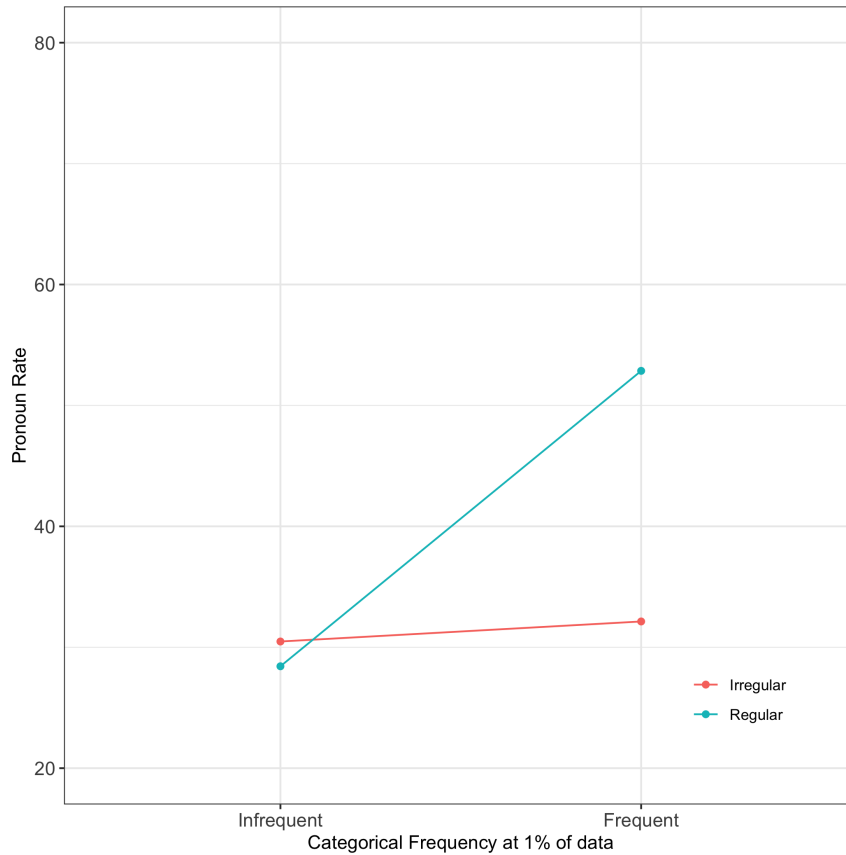


Figure 4.7: Morphological regularity: frequent vs. infrequent forms

(52.9% and 28.4%, respectively). Although small, there is also an increase in pronoun rate from infrequent (30.5%) to frequent irregular forms (32.1%).

Two mixed effects logistic regression models were run to determine the extent to which adding an interaction term produces a significantly more explanatory model. Interaction terms such as the one included here are meant to capture possible synergies between two variables, i.e. a combined effect that is different from separate individual effects. The first model included morphological regularity and Discrete Frequency as main effects. The second model includes Discrete Frequency and morphological regularity in interaction. Both models include verb as a random effect.

The second of the two models, the interaction model, is presented in Table 4.5. Model results indicate that there is a significant interaction between Morphological

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.72	0.25	-2.86	0.004
Regular	0.89	0.36	2.44	0.015*
Infrequent	-0.42	0.26	-1.64	0.1
Regular*Infrequent	-0.96	0.37	-2.63	0.009**

Table 4.5: Interaction between Morphological Regularity and Discrete Frequency; Model configuration: `glmer(Pronoun ~ Regularity*Discrete Frequency)`; Reference levels: Irregular, Frequent, Irregular*Frequent

Regularity and Frequency ($\beta = -0.96, p = 0.009$). One surprising result is that there is no significant main effect of discrete frequency ($p = 0.1$). At the very least, this result casts some doubt on the nature of the frequency effect reported in Section 4.2. The marginal R^2 for the interaction model is 0.035, and the conditional R^2 is 0.221. An ANOVA was run to compare the two models, the one with and without an interaction term, respectively. ANOVA results reveal that these models are significantly different, and the interaction model is more explanatory, since its AIC is 2 lower than the AIC of the model with no interaction term ($p = 0.035$).

Taken together, these results show that lexical frequency is potentiating an effect of morphological regularity. An effect of morphological regularity is clearly enhanced for frequent forms: frequent regular verbs have a pronoun rate that is more than 20% higher than frequent irregular forms. In contrast, infrequent forms hover around 30%, regardless of morphological regularity. Like the morphological regularity results in Erker and Guy (2012), these findings point to an activation of a linguistic constraint via interaction with frequency. The genuine evidence for a relationship between the (ir)regularity of verbal inflectional morphology and the probability of subject pronoun use is, upon closer inspection, largely restricted to frequently occurring verb forms.

4.4 Person and Number

Pronoun rates significantly differ across different person and number combinations. Table 4.6 displays the raw counts and pronoun rates for all Person/Number combinations organized by pronoun rate in the data. Third-person singular forms have the highest overt pronoun rate at 39.8%. Second-person singular has the second highest pronoun rate at 36.4%, followed by first-person singular (35.7%). Third-person plural forms have an overt pronoun rate of 15.3%, which is followed by first-person plural (12.4%) and second-person plural (5.9%).

	<i>N</i> VERBS	% OVERT PRONOUNS
3rd singular	16,325	39.8
2nd singular	8,631	36.4
1st singular	42,828	35.7
3rd plural	13,048	15.3
1st plural	7,105	12.4
2nd plural	34	5.9

Table 4.6: Person and number of the verb, all combinations considered

A mixed effects logistic regression model that includes PERSON/NUMBER as a fixed effect and VERB as a random effect was run to assess the explanatory power of Person/Number on pronoun production. Second person plural was excluded from the model due to its low number of observations. Model results (presented in Table 4.7) show that all levels of Person/Number significantly impact pronoun production ($ps < 0.05$). First, second, and third singular are all significantly more likely to occur with a pronoun compared to the reference level ($ps < 0.001$). Third plural is also significantly more likely to occur with a pronoun than the reference level, though the effect is not a strong ($\beta = 0.188, p = 0.021$). The marginal R^2 is 0.069, and

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.115	0.062	-34.092	< 0.001***
First Singular	1.392	0.067	20.864	< 0.001***
Second Singular	0.977	0.079	12.436	< 0.001***
Third Plural	0.188	0.075	2.499	0.012*
Third Singular	1.387	0.068	20.290	< 0.001***

Table 4.7: Main effect of Person/Number; Model configuration: `glmer(Pronoun ~ Person/Number + (1 | Verb))`; Reference level: First Plural

the conditional R^2 is 0.186. These statistical results are generally consistent with those reported in Erker and Guy (2012). However, their individual rates were slightly different: second-person singular had the highest pronoun rate at 48% in their study, which is much higher than is reported in this dissertation.

4.4.1 Interaction with Discrete Frequency

Similar to morphological regularity, the results for Person/Number reveal an interaction effect of Discrete Frequency. First, we will look at the singular forms in isolation (in following Erker and Guy 2012). Figure 4.8 shows the pronoun rates by Discrete Frequency for first person, second person, and third person singular forms. There are large differences in the range of pronoun rates for frequent forms (from 77% to 36.7%). In contrast, the pronoun rates for infrequent forms are confined to a range of roughly 10%, with second singular forms at the bottom end of that range (29.6%) and third singular forms at the top (38.8%). Importantly, the extremely high range of pronoun use for frequent forms is effectively due to *sabes* alone, since it is the only second person singular form labelled “frequent” in this corpus. Nevertheless, the dif-

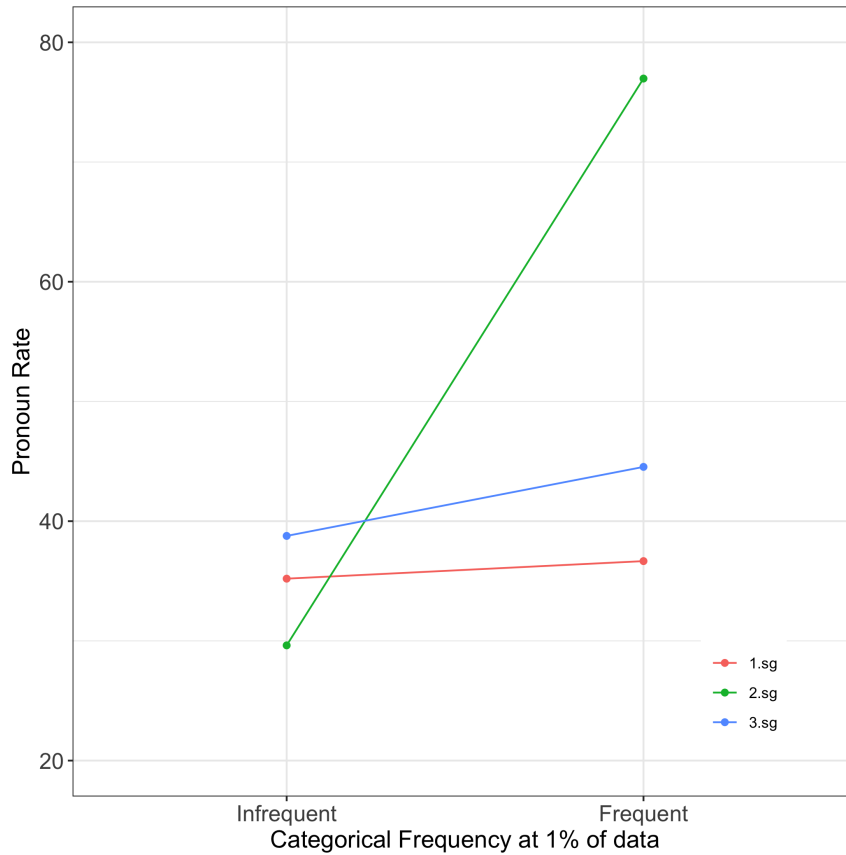


Figure 4-8: Subset of Person/Number: frequent vs. infrequent forms

ference between first and third person singular, which are based on multiple forms, is amplified in frequent verbs (Range difference = 28.6).

When we consider all six Person/Number combinations, this trend persists. Figure 4-9 shows the pronoun rates for all Person/Number combinations for frequent and infrequent verb forms. First person plural and second person plural forms are only represented in the infrequent category because there are no frequent forms with these person/number factor values. This illustration reinforces what was presented in Figure 4-8. We see a significantly larger range of pronoun rates for frequent verb forms ($range = 61.5$) than for infrequent verb forms ($range = 32.9$). Once again, the present findings support the notion that Discrete Frequency amplifies the impact of other linguistic constraints on variable pronoun production.

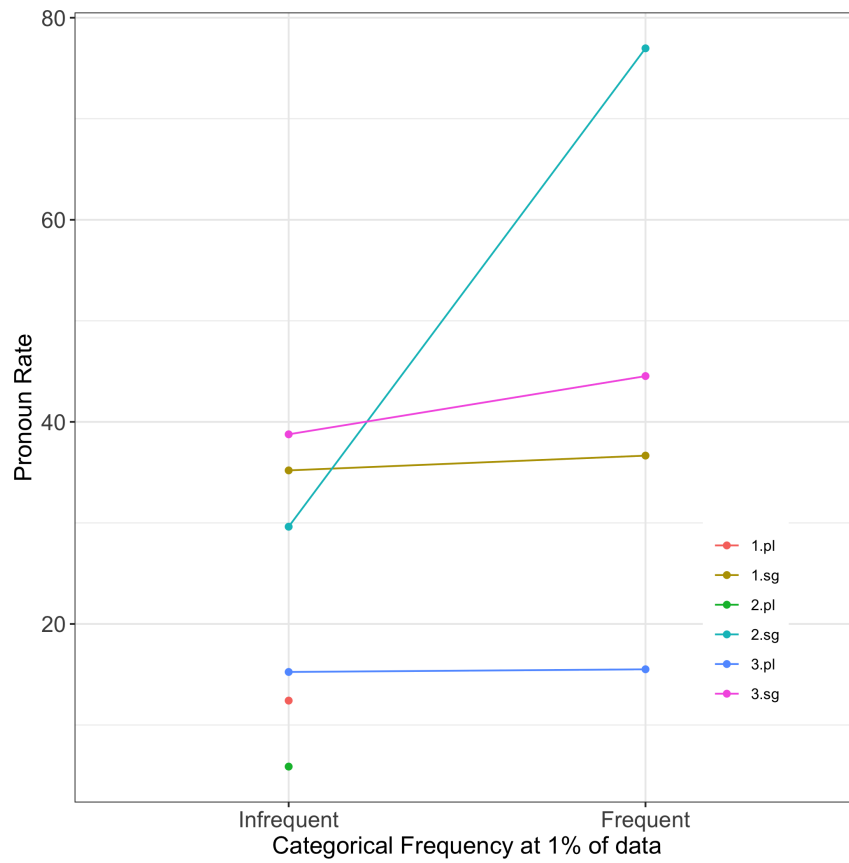


Figure 4·9: Person/Number: frequent vs. infrequent forms, all forms considered

To explore the potential of an interaction, and whether such an interaction amounts to an increase in explanatory power, two mixed effects logistic regression models were run. The first mixed effects model includes Person/Number and Discrete Frequency as fixed effects and verb as a random effect. The second model includes Person/Number and Discrete Frequency as an interaction and verb as a random effect. Second person plural was excluded in both models. The interaction model output is presented in Table 4.8. There is a significant interaction between second person singular and infrequent forms ($p < 0.001$). Interestingly, the main effect of discrete frequency is not significant ($p = 0.174$). It is important to underscore the lack of significance for Discrete Frequency in this interaction model as well as the interaction model for morphological regularity (Table 4.5). Taken together with the wide ranging variation in rates at the upper end of the frequency spectrum (Figure 4.6), these results weaken the case for a monotonic effect of frequency.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.841	0.208	-8.842	< 0.001***
First Singular	1.385	0.105	13.172	< 0.001***
Second Singular	2.133	0.329	6.486	< 0.001***
Third Plural	0.722	0.336	2.152	0.031*
Third Singular	1.388	0.068	20.454	< 0.001***
Infrequent	-0.272	0.200	-1.360	0.174
First Singular*Infrequent	0.007	0.088	0.077	0.938
Second Singular*Infrequent	-1.182	0.327	-3.611	< 0.001***
Third Plural*Infrequent	-0.542	0.334	-1.622	0.105

Table 4.8: Person/Number & Discrete Frequency Interaction; Model configuration: `glmer(Pronoun ~ Person/Number*Discrete Frequency + (1 | Verb))`; Reference levels: First plural, Frequent, First plural*Frequent

Results of an ANOVA comparing these two models reveal that, once again, the regular model and the interaction model are significantly different. And, the interaction model is more explanatory than the regular model, with an AIC that is 7 points lower than the regular model (Regular model $AIC = 101,487$; Interaction model $AIC = 101,480$, $p = 0.003$). Although AICs reveal the interaction model is stronger, R^2 measures tell a somewhat different story: The marginal and conditional R^2 are slightly lower for the interaction model ($R_m^2 = 0.084$; $R_c^2 = 0.197$) than the regular model ($R_m^2 = 0.086$; $R_c^2 = 0.200$). Nevertheless, these general trends are comparable to Erker and Guy (2012), who also found a large difference in pronoun rates across frequencies and report a similarly large jump in pronoun rate for second-person singular in particular.

4.5 TMA

Table 4.9 shows all possible levels of TMA along with their N and percent of overt SPP. Here we see quite a bit of variation in pronoun production for different TMA levels. Perfect subjunctive has the highest pronoun rate at 60%. In contrast, imperative forms have the lowest pronoun rate at 3.8%, which is to be expected because imperative forms are typically used without subject pronouns. It is interesting that the Conditional has the second highest SPP rate (37.4%) since it had the second *lowest* pronoun rate in Erker and Guy (2012) (17%). Imperfect indicative verbs occurred with SPPs most frequently in Erker and Guy (2012) (43%), but have the third highest rate (37.2%) in the present study. The periphrastic future has the third lowest SPP rate of all TMA at 25.1%. This rate is 8.5% lower than the pronoun rate for present indicative, which goes against expectations since the periphrastic future consists of a present indicative form followed by an infinitival phrase. The perfect indicative, which also contains a present indicative inflected form, has an SPP rate that is much

closer to present indicative (30.3%).

	<i>N</i> VERBS	% OVERT PRONOUNS
Perfect subjunctive	10	60
Conditional	1,034	37.4
Imperfect indicative	12,712	37.2
Present indicative	47,375	33.6
Past subjunctive	935	32.7
Perfect indicative	3,217	30.3
Preterite indicative	16,221	26.3
Present Subjunctive	2,275	26
Periphrastic future	2,045	25.1
Future indicative	188	24
Imperative	1,988	3.8

Table 4.9: TMA of the verb, all combinations considered

Along with variation in pronoun rates across TMA levels, Table 4.9 also reveals variation in the number of observations for each TMA level. For example, the present indicative appears 47,375 times in the corpus (constituting over half of all tokens in the dataset), but the perfect subjunctive appears only 10 times. Because of these asymmetries in *N*s, TMA is a particularly challenging variable to analyze statistically. I will deal with this issue in the following ways: (1) collapsing perfect subjunctive and future indicative forms into a single category labelled ‘other’; (2) running statistical models on the full dataset with all TMA levels; (3) running statistical models that include only the top three TMA categories (present indicative, preterite indicative, and imperfect indicative).

A mixed effects logistic regression model was run to examine the relationship between TMA and subject pronoun production. TMA was included as a fixed effect

and verb was included as a random effect. Model results are presented in Table 4.10.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.21	0.12	-9.75	< 0.001***
Other	0.07	0.23	0.28	0.77
Imperative	-2.33	0.18	-12.41	< 0.001***
Imperfect Indicative	0.21	0.13	1.59	0.11
Past Subjunctive	0.12	0.16	0.73	0.46
Perfect	0.15	0.14	1.09	0.28
Periphrastic Future	-0.04	0.15	-0.29	0.77
Present Indicative	0.26	0.13	2.04	0.04*
Present Subjunctive	-0.19	0.15	-1.32	0.19
Preterite Indicative	-0.14	0.13	-1.08	0.28

Table 4.10: Main effect of TMA; Model configuration: `glmer(Pronoun ~TMA + (1 | Verb))`; Reference level: Conditional

Results show a statistically significant effect for certain levels of TMA. Specifically, imperative forms are significantly less likely to occur with a pronoun ($\beta = -2.33, p < 0.001$) compared to the reference level conditional. Present indicative forms are significantly more likely to occur with a pronoun ($\beta = 0.26, p = 0.04$). All other TMA levels are not statistically significantly different when contrasted with the reference level. The marginal R^2 is 0.039 and the conditional R^2 is 0.206 for this model.

Due to the very large discrepancy in the Ns for the factor values of TMA, a second main effects model was run that included only the three most frequent factor values (imperfect indicative, present indicative, and preterite indicative). As with the full TMA model, verb was included as a random effect. Results of this mixed effects

logistic regression model are presented in Table 4.11.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.00	0.05	-21.20	< 0.001***
Present Indicative	0.05	0.06	0.95	0.34
Preterite Indicative	-0.36	0.06	-6.03	< 0.001***

Table 4.11: Main effect of top three factor values of TMA; Model configuration: `glmer(Pronoun ~ TMA + (1 | Verb))`; Reference level: Imperfect indicative

Model results show that only the preterite indicative level is significantly different from the reference level imperfect indicative. Preterite indicative forms are less likely to occur with a pronoun ($\beta = -0.36, p < 0.001$). The marginal R^2 for this model is 0.007 and the conditional R^2 is 0.181.

4.5.1 Interaction with Discrete Frequency

Figure 4-10 plots the pronoun rates separated by Discrete Frequency for three TMA levels (imperfect indicative, present indicative, and preterite indicative). These levels were chosen because they are the three most-frequent, and they correspond to the three TMA levels that contained frequent and infrequent forms that Erker and Guy (2012) could reasonably compare. As the figure shows, pronoun rates for frequent forms are higher than pronoun rates for infrequent forms. Interestingly, however, the frequency effects depicted in the figure suggest the absence of evidence for an interaction, in contrast to Erker and Guy (2012). Frequent verb forms, with a range of 6.4% (36.1% to 42.5%) present smaller amounts of pronoun rate variation than infrequent forms ($range = 10.5\%; 26.3\% \text{ to } 36.8\%$). This is inconsistent with the notion that high frequency magnifies the effect of other linguistic predictors, which would predict an increased range between frequent forms. Instead, we see a higher

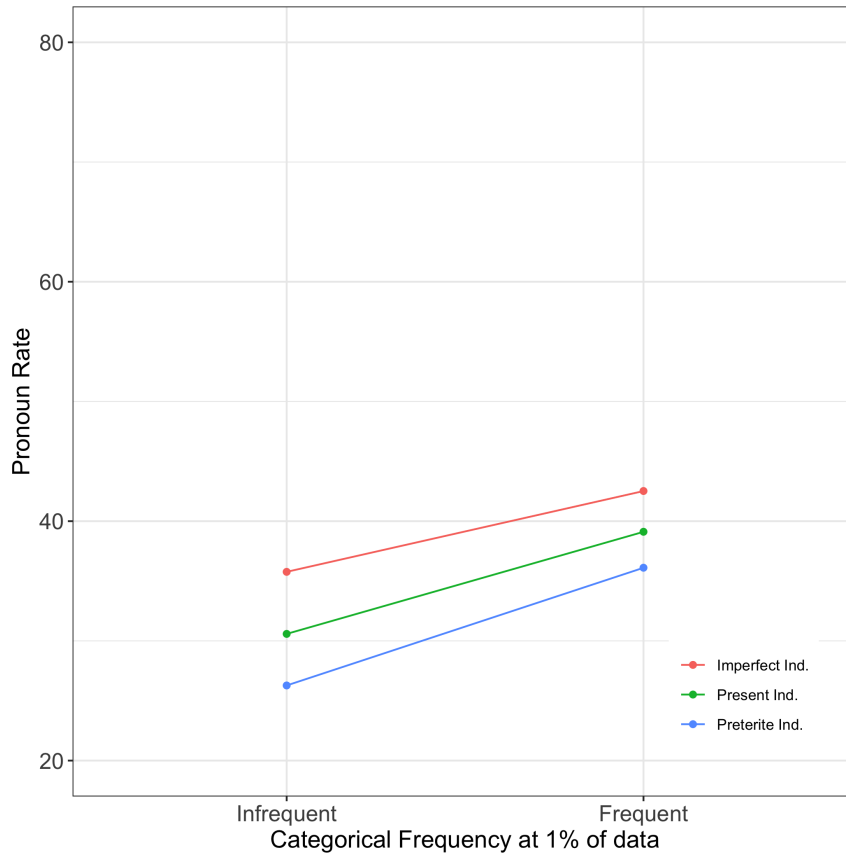


Figure 4·10: Subset of TMA: frequent vs. infrequent forms

range in pronoun rates between TMA levels for infrequent forms.

In order to test the statistical significance of the trends depicted in Figure 4·10, two mixed effects logistic regression models were run on the subset of TMA data. The first model includes TMA and Discrete Frequency as fixed effects and verb as a random effect. The second model includes TMA and Discrete Frequency as an interaction and verb as a random effect. Results, which are shown in Table 4.12, reveal no significant effects for any level in the model. The marginal R^2 for this model is also quite low, and it is identical to the that of the regular model ($R_m^2 = 0.025$).

Further, upon running an ANOVA to compare the regular and interaction model, results indicate that these two models are not significantly different from one another ($p = 0.52$). These results indicate that, at least for this subset of TMA, there is no

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.35	0.34	-1.01	0.31
Present Indicative	-0.09	0.34	-0.27	0.78
Preterite Indicative	-0.06	0.42	-0.15	0.88
Infrequent	-0.66	0.35	-1.90	0.06
Present Ind.*Infrequent	0.14	0.35	0.42	0.68
Preterite Ind.*Infrequent	-0.29	0.42	-0.70	0.48

Table 4.12: Interaction between a subset of TMA & Discrete Frequency; Model configuration: `glm(Pronoun ~TMA*Discrete Frequency)`; Reference levels: Imperfect indicative, Frequent, Imperfect indicative*Frequent

interaction between TMA and Discrete Frequency. Due to these opaque results, we will now turn to an expanded analysis of TMA that includes all factor values in the full dataset (with the exception of “other” which collapses perfect subjunctive and future indicative).

Two standard logistic regression models⁶ were run on the full dataset to test the potential for an interaction. The first model included TMA and Discrete Frequency as fixed effects and pronoun presence/absence as the dependent variable. The second model included TMA and Discrete Frequency as an interaction.

The model output for the full interaction model is presented in Table 4.13. Model results indicate that there is no significant contrast between Infrequent and Frequent forms and none of the interaction levels returned significant results. The Tjur’s R^2

⁶As with previous interactions, two mixed effects logistic regression models were run (one including TMA and Discrete Frequency as main effects and the other including an interaction). The interaction model failed to converge, likely due to the extreme disparity in frequencies of frequent and infrequent verb forms, and their connection to verb.

⁷The Tjur’s R^2 is reported because this model is a fixed effects model, which means it does not account for any random effects.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.05	0.35	-0.15	0.88
Other	-0.54	0.17	-3.12	0.002**
Imperative	-2.73	0.13	-20.33	< 0.001***
Imperfect Indicative	-0.25	0.36	-0.71	0.48
Past Subjunctive	-0.21	0.09	-2.18	0.03*
Perfect	-0.32	0.07	-4.24	< 0.001***
Periphrastic Future	-0.74	0.36	-2.03	0.04*
Present Indicative	-0.39	0.35	-1.10	0.27
Present Subjunctive	-0.53	0.08	-6.66	< 0.001***
Preterite Indicative	-0.52	0.07	-7.76	< 0.001***
Infrequent	-0.46	0.35	-1.33	0.18
Imperfect Ind.*Infrequent	0.18	0.35	0.51	0.61
Periphrastic Fut.*Infrequent	-0.04	0.36	-0.10	0.92
Present Ind.*Infrequent	0.08	0.35	0.24	0.81

Table 4.13: Interaction between TMA & Discrete Frequency; Model configuration: `glm(Pronoun ~TMA*Discrete Frequency)`; Reference levels: Conditional, Frequent, Conditional*Frequent

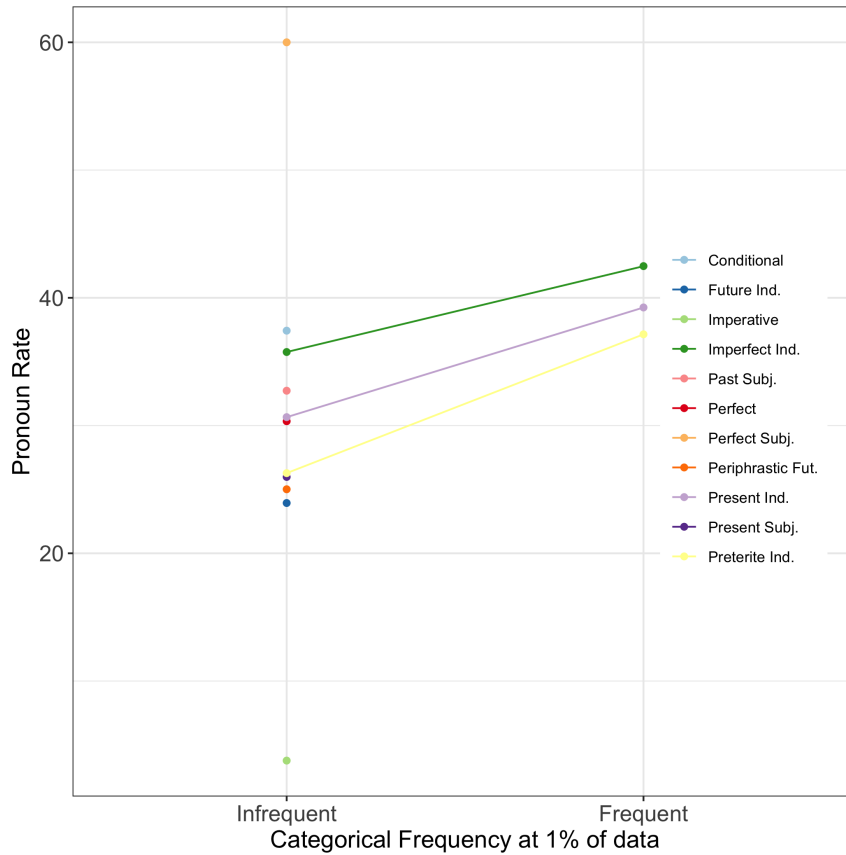


Figure 4-11: TMA: frequent vs. infrequent forms, all forms considered

for this model is 0.020. An ANOVA was run to compare the interaction model with the main effect model. The AIC is 6 points lower for the interaction model ($AIC = 107,812$), than the regular model ($AIC = 107,818$). However, ANOVA results reveal that these models are not significantly different ($p = 0.12$), suggesting that there is not an interaction between Discrete Frequency and TMA.

Erker and Guy (2012) do not provide figures that include all levels of TMA. Nevertheless, given the unusual result presented in Figure 4-10, the full figure is presented here (see Figure 4-11). When considering all levels, we still see an opposite effect than presented in Erker and Guy (2012). There is more variation between TMA levels for infrequent forms than for frequent forms. Pronoun rates are split such that the infrequent forms generally hover between 20% and 40%, while the

frequent forms range from 30% up to 42.5%. Of course, it bears mention that there are fewer TMA levels in the frequent group, since not all TMA factor values occurred at a rate of at least 1%, and this is likely contributing to the model results presented here. Nevertheless, one would expect that TMA would be closely tied to the verb and therefore one would predict a significant interaction between TMA and Discrete Frequency.

4.6 Semantic Category

For semantic content of the verb, results indicate that mental activity verbs correspond to the highest overt pronoun rates (39.2%). Stative verbs have the second highest pronoun rate (32.4%), and External activity verbs have the lowest rate (28.6%). These rates (shown in Table 4.14) are consistent with what was found in Erker and Guy (2012): Mental activity verbs had the highest overt pronoun rates (45%), then stative verbs (36%), and finally, external activity verbs (31%). Again, we see lower overall pronoun rates in the present corpus than in Erker and Guy (2012).

	<i>N</i> VERBS	% OVERT PRONOUNS
Mental	16,595	39.2
Stative	23,380	32.4
External	48,026	28.6

Table 4.14: Semantic content of the verb

A mixed effects logistic regression model was run to investigate the extent to which semantic category impacts pronoun production. Semantic category was included as a fixed effect and verb was included as a random effect.

Table 4.15 displays the model outputs. Results reveal that both mental activity verbs ($\beta = 0.13$) and stative verbs ($\beta = 0.12$) are significantly more likely to occur

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.22	0.02	-54.04	$p < 0.001$ ***
Mental Activity	0.13	0.04	3.68	$p < 0.001$ ***
Stative	0.12	0.03	3.66	$p < 0.001$ ***

Table 4.15: Main effect of semantic category; Model configuration: `glmer(Pronoun ~ Semantic category + (1 | Verb))`; Reference level: External activity verbs

with a pronoun than the reference level, external activity verbs ($ps < 0.001$). Crucially, though, the random effect of verb is doing the heavy lifting in this model, since the Marginal R^2 is only 0.001 and the conditional R^2 is 0.193.

4.6.1 Interaction with Discrete Frequency

Figure 4.12 plots the pronoun rates for frequent and infrequent forms based on their semantic category. Overall, there are higher pronoun rates for frequent forms than infrequent forms, and the different semantic classes present pronoun rates in line with previous literature – with pronoun rates highest for mental activity verbs and lowest for external activity verbs. The results in Erker and Guy’s study show extreme differences in the pronoun rates between infrequent and frequent forms. They attribute this to an activation effect of frequency. Different from an amplification effect, which increases an existing effect for frequent forms, activation essentially *turns on* an effect such that it only occurs for frequent forms. According to Erker and Guy (2012), this activation effect explains the lack of significant difference between pronoun rates across semantic categories for infrequent verb forms. In the present study, the differences between SPP rates for frequent and infrequent forms in terms of semantic category is not as robust. Although not identical, these results are in-line with Erker and Guy (2012). There is a clear frequency effect: frequency in this corpus is acting as an amplifier, not an activator. We see a clear effect of semantic category for infre-

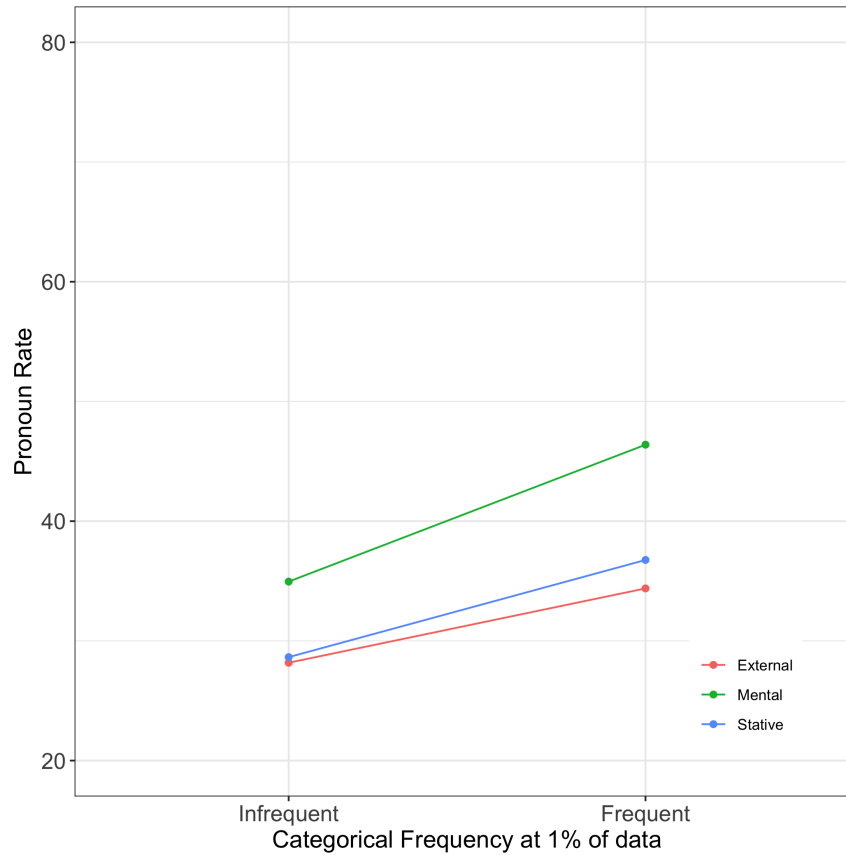


Figure 4.12: Semantic content: frequent vs. infrequent forms

quent forms, but an increase in variation of pronoun rates when we look at frequent forms. Pronoun rates range from 34.4% to 46.4% ($Range = 12$) for frequent forms and 28.2% to 35% ($Range = 6.8$) for infrequent forms.

ANOVA results from Erker and Guy (2012) show no significant difference in pronoun rates across semantic category for infrequent forms, but do corroborate the pronoun rate differences across semantic class for frequent forms. In the current study, ANOVA results comparing two mixed effects logistic regression models (one with an interaction between Switch Reference and Discrete Frequency and one with only a main effect) indicate that the two models are significantly different. Model results for the interaction model are presented in Table 4.16. Model results show significant interactions for the Mental*Infrequent level ($p < 0.001$) and Stative*Infrequent level

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.35	0.26	-1.39	0.164
Mental	-0.25	0.10	-2.43	0.015*
Stative	-0.07	0.08	-0.88	0.38
Infrequent	-0.88	0.26	-3.42	0.001 ***
Mental*Infrequent	0.44	0.11	4.03	< 0.001 * **
Stative*Infrequent	0.18	0.09	2.02	0.04*

Table 4.16: Interaction between Semantic Category and Discrete Frequency; Model configuration: `glmer(Pronoun ~ Semantic category*Discrete Frequency + (1 | Verb))`; Reference levels: External activity verb, Frequent, External activity verb*Frequent

($p = 0.04$). The interaction model also shows a significant main effect of discrete frequency ($p = 0.001$). We see a large increase in R^2 for the interaction model compared to that of the model that only included Semantic Category (Table 4.15): Marginal R^2 for the interaction model is 0.023 (compared to 0.001 for the main effect model), and the Conditional R^2 is 0.210. Further, the interaction model is more explanatory with an AIC that is 14 points lower than the regular model ($p < 0.001$). This finding suggests an interaction between semantic category and Discrete Frequency such that the effects of semantic category on pronoun production are heightened for frequent forms. These findings support the amplification effect of frequency outlined in Erker and Guy (2012).

4.7 Switch Reference

Results for switch reference mirror those of previous work that has investigated this variable as a main effect of pronoun variation. Tokens that appear with a switch in

referent correspond to higher overt pronoun rates (37.7%) than tokens that appear in contexts without a switch in referent (24.6%).

	<i>N</i> VERBS	% OVERT PRONOUNS
Switch in referent	47,019	37.7
Same referent	40,982	24.6

Table 4.17: Switch Reference

A mixed effects logistic regression model with switch reference as a fixed effect and verb as a random effect was run to determine the impact of switch reference on pronoun production. The results of this model are presented in Table 4.18.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.89	0.02	-37.97	< 0.001 * **
Same	-0.76	0.02	-45.94	< 0.001 * **

Table 4.18: Main effect of Switch Reference; Model configuration: `glmer(Pronoun ~ Switch Reference + (1 | Verb))`; Reference level: Different referent

Model results show that a phrase with the same referent is significantly less likely to occur with an overt pronoun ($\beta = -0.76, p < 0.001$). The marginal R^2 is 0.033, while the conditional R^2 is 0.242. Erker and Guy (2012) find a similar effect: there is a significant difference between pronoun rates for switch (40%) and no switch contexts (29%) in their study.

4.7.1 Interaction with Discrete Frequency

Figure 4-13 plots pronoun rates for switch reference based on Discrete Frequency. The findings here are strikingly similar to the findings in Erker and Guy (2012). We see a larger effect of switch reference for frequent verb forms than we do for infrequent

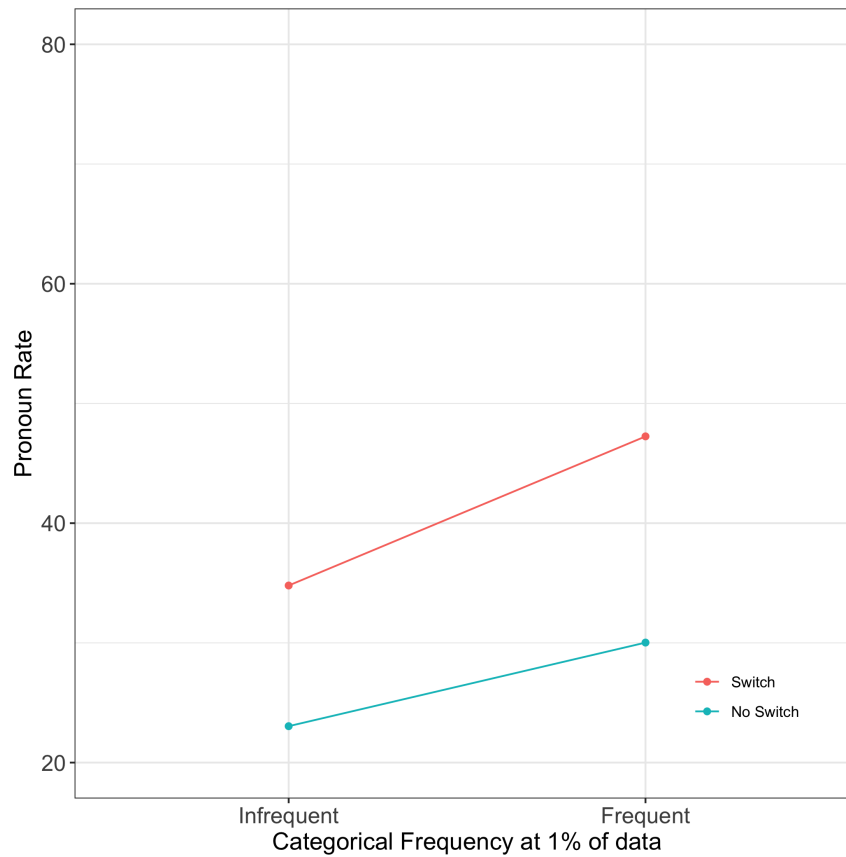


Figure 4.13: Switch reference: frequent vs. infrequent forms

verb forms. That is to say that there is greater disparity between the pronoun rates of switch and no switch when the verb is frequent (*Switch* = 47.4, *Noswitch* = 30.3; *Range* = 17.1) than when the verb is infrequent (*Switch* = 34.9, *Noswitch* = 23.1; *Range* = 11.8).

As with the previous predictor variables, two mixed effects logistic regression models were run. One model included switch reference and Discrete Frequency as individual fixed effects. The other model included these two predictors as main effects *and* as an interaction. Both models included verb as a random effect. The model output for the interaction model is presented in Table 4.19. The marginal R^2 for this model is 0.059, and the conditional R^2 is 0.264. Model results show a significant main effect of discrete frequency ($p = 0.007$). There is also a significant interaction for the

Same Referent*Infrequent level ($p < 0.001$).

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.17	0.26	-0.63	0.53
Same Referent	-0.62	0.03	-19.45	< 0.001 * **
Infrequent	-0.71	0.26	-2.7	0.007**
Same Referent*Infrequent	-0.2	0.04	-5.28	< 0.001 * **

Table 4.19: Interaction between Switch Reference and Discrete Frequency; Model configuration: `glmer(Pronoun ~ Switch Reference*Discrete Frequency + (1 | Verb))`; Reference levels: different referent, frequent, different referent*frequent

After running these models, an ANOVA was then run to determine the extent to which these models differ. ANOVA results show that the two models are significantly different, and the interaction model has more explanatory power with an AIC that is 26 points lower than the AIC for the regular model ($p < 0.001$). These findings provide stronger support for the role of frequency than those found in Erker and Guy (2012), since they only found a near-significant interaction between switch reference and Discrete Frequency ($p = 0.057$).

4.8 Summary of Replication Results

Table 4.20 presents a summary of the main findings from the analyses presented above, including main effects, interaction effects, and amplification effects. To determine the extent to which each linguistic constraint has a significant main effect on pronoun use, ANOVAs were run comparing each main effect model (including the linguistic constraint as a main effect and verb as a random effect⁸) to the null model (including

⁸Because the TMA mixed effects regression models failed to converge, standard logistic regression models were used for model comparisons to determine a significant main effect of TMA.

only verb as a random effect). The p-values reported in the “main effect” column are taken directly from these ANOVA results.

CONDITIONING FACTOR	MAIN EFFECT	INTERACTION W/FREQUENCY	EVIDENCE OF AMPLIFICATION EFFECT
Morph Regularity	YES $p = 0.045$	YES $p = 0.035$	YES Range diff. = 18.7 (larger in freq. forms)
Person/Number	YES $p < 0.05$	YES $p = 0.003$	YES Range diff. = 28.6
TMA	YES $p < 0.001$	NO $p = 0.12$	NO Range diff. = -4.1
Semantic content	YES $p < 0.001$	YES $p < 0.001$	YES Range diff. = 5.2
Switch reference	YES $p < 0.001$	YES $p < 0.001$	YES Range diff. = 5.5

Table 4.20: Summary of main effects, interactions with frequency, and evidence of amplification effects for core constraints

Multivariate analyses reveal significant main effects for all linguistic constraints. Results also show significant interactions between the variable Discrete Frequency and each linguistic constraint except TMA. Finally, the last column in the table reports evidence of an amplification effect using the difference in range between frequent and infrequent forms for each linguistic predictor. As the table shows, all predictors with the exception of TMA show an amplification effect of frequency such that the range of pronoun rates for frequent forms is greater than the range in pronoun rates for infrequent forms. These findings provide some of the strongest support for the notion of an amplification effect of frequency since Erker and Guy (2012). The results do stray slightly from the findings in Erker and Guy (2012), who report no main effect of morphological regularity, semantic content, or Person/Number on infrequent forms. The interaction results are generally consistent with Erker and Guy (2012), who find a near significant interaction between TMA and Discrete Frequency. The current study finds a significant interaction between switch reference and frequency, which is not borne out in Erker and Guy (2012). As previously mentioned, the multivariate analyses carried out here differ from the univariate analyses carried out in Erker and Guy (2012). These differences are not large enough that they draw the current findings into question.

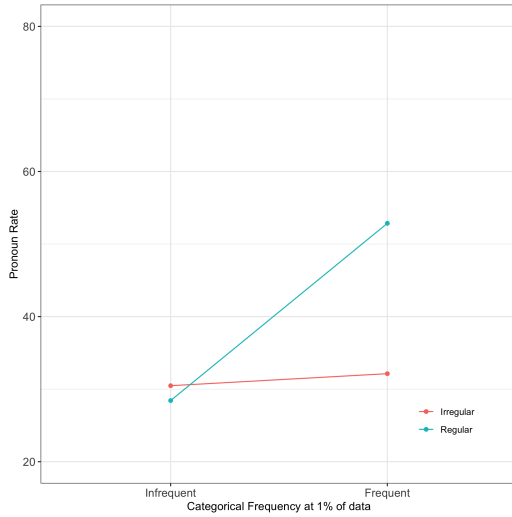
To better understand the extent to which each of these linguistic predictors constrain variable pronoun use, Table 4.21 presents the R^2 values for all main effects and interactions⁹. In each case, the inclusion of a frequency-based interaction term increases each models' Marginal R^2 , thereby increasing each model's strength and ability to account for the variation observed in the data. The linguistic constraint with the highest marginal R^2 for the main effects models and the interaction models is Person/Number. This means that including Person/Number as the only predictor in the model accounts for 6.9% of the observed variation in the data, and including it as an interaction with Discrete Frequency accounts for 8.4% of the variation in the data. Taken together, Tables 4.20 and 4.21 show that overall frequency (presented here as discrete frequency) interacts with other linguistic predictors in order to increase their effects on pronoun use.

CONDITIONING FACTOR	MAIN EFFECT	INTERACTION W/FREQUENCY
Morph Regularity	$R_m^2 = 0.00$; $R_c^2 = 0.195$	$R_m^2 = 0.035$; $R_c^2 = 0.221$
Person/Number	$R_m^2 = 0.069$; $R_c^2 = 0.186$	$R_m^2 = 0.084$; $R_c^2 = 0.197$
TMA	$R_m^2 = 0.039$; $R_c^2 = 0.206$	
Semantic content	$R_m^2 = 0.001$; $R_c^2 = 0.193$	$R_m^2 = 0.023$; $R_c^2 = 0.210$
Switch reference	$R_m^2 = 0.033$; $R_c^2 = 0.242$	$R_m^2 = 0.059$; $R_c^2 = 0.264$

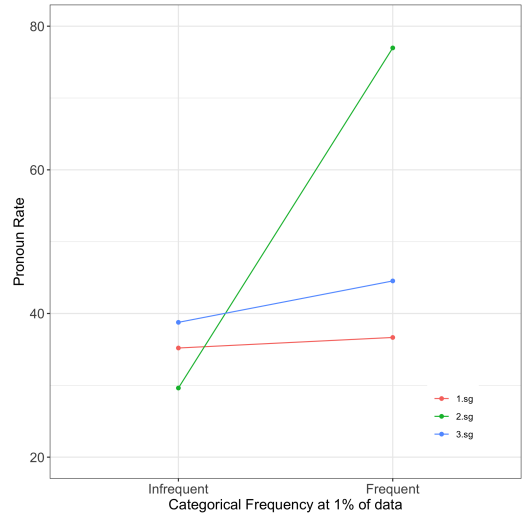
Table 4.21: Summary of marginal and conditional R^2 for the main effect and interaction models of core constraints

At first glance, this study, unlike Erker and Guy (2012), seemingly provides much stronger statistical support for a direct effect of verb frequency on pronoun production, i.e. that higher frequency significantly increases the likelihood of pronoun use. The significant positive correlations for raw and log frequency and the observed 10%

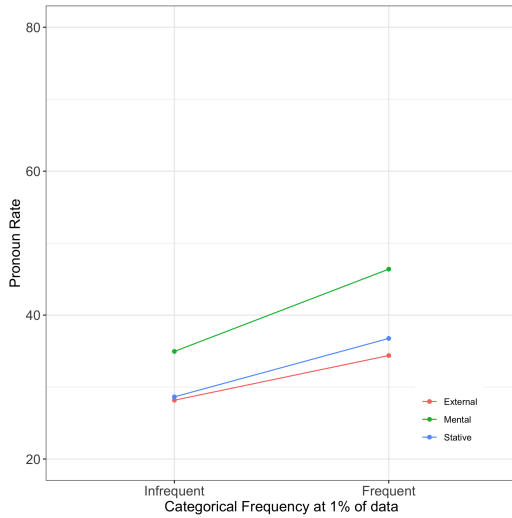
⁹The R^2 for the full TMA/Discrete Frequency interaction model is excluded because a fixed effects model was run, which corresponds to a Tjur R^2 that is not directly comparable to the marginal and conditional R^2 s for the other models.



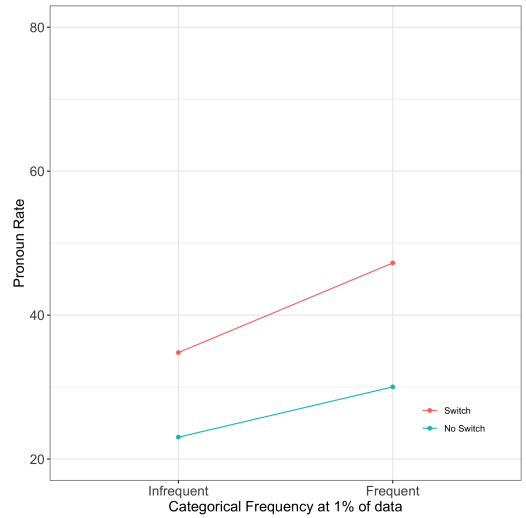
(a) Morphological Regularity x Discrete Frequency



(b) Person/Number x Discrete Frequency



(c) Semantic Content x Discrete Frequency



(d) Switch Reference x Discrete Frequency

Figure 4.14: Amplification effects of Discrete Frequency

increase in pronoun rates for frequent forms compared to infrequent forms suggest this. However, deeper investigation underscores an opacity in the relationship between lexical frequency and pronoun use that corroborates the findings in Erker and Guy (2012). The immense variation in pronoun rates presented in Figures 4·2 and 4·3 and the inconsistent pronoun rates for the 32 most frequent verbs forms provide sufficient evidence *against* a monotonic effect of frequency on variable pronoun production. The present findings support the notion that frequency effects work “under the hood” to either amplify or activate linguistic predictor effects. Frequency is not simply impacting pronoun use directly (at least for forms with a log frequency below 4, and otherwise with very many exceptions), but it is indirectly impacting pronoun use by interacting with other linguistic constraints. This is especially evident when we compare all interaction figures side by side¹⁰ (see Figure 4·14). The figure shows that the effects of linguistic constraints are more robust for frequent finite forms, which suggests that mental representations of high-frequency forms include increased sensitivity to other linguistic predictors that is not as robust in the representations of low-frequency forms. This is unsurprising given that high-frequency verb forms correspond to verbs that speakers encounter (through speaking or listening) most often. Per UBG, this increase in exposure thereby increases the details of these verbs’ mental representations, which ultimately leads to sensitivity to other linguistic constraints.

The following section investigates potentially better-suited definitions for two linguistic predictors from Erker and Guy (2012): Morphological Regularity 2.0, and Semantic Category 2.0. In addition, the following section expands upon the current analysis in an investigation of Preceding Pronoun. The investigation into these three constraints aims to shed light on the nature of these variables as constraints on pronoun use and to determine their interaction –if any– with discrete frequency.

¹⁰TMA was excluded from this comparison since there was no significant interaction between TMA and Discrete Frequency.

4.9 Exploring Additional Linguistic Constraints

Before analyzing the contextual frequency metrics, we must first investigate the main effects of three other linguistic predictors that were not investigated in Erker and Guy (2012): PRECEDING PRONOUN, MORPHOLOGICAL REGULARITY 2.0, and SEMANTIC CATEGORY 2.0. Morphological regularity as a variable *was* investigated in Erker and Guy (2012) as a binary predictor that categorized verbs as either regular or irregular. The results of that variable are presented in Section 4.3. However, the present study proposes a new way to consider Morphological regularity that is more consistent with verbal morphology research in the field of morphophonology (Albright et al., 2000; Albright and Hayes, 2003). The results for this variable and for the preceding pronoun are presented in the next two sections. Semantic category was also investigated in Erker and Guy (2012), and the replication results are described in Section 4.6. However, as discussed in Chapter (1), more recent work on semantic category has drawn this variable into question. In an effort to best understand contextual frequency as it relates to other constraints, the present study will also investigate semantic category as a variable with four levels (referred to as Semantic Category 2.0): mental activity, stative, estimative, and external activity. The results for this analysis is presented in Section 4.9.3.

4.9.1 Preceding Pronoun

Table 4.22 presents the pronoun rates for tokens with a pronoun in the preceding site of pronominal variation and without a pronoun in the preceding site of pronominal variation. As the table shows, there seems to be a clear effect of Preceding Pronoun such that tokens with a preceding pronoun favor overt pronoun production (44.2%) and contexts without a preceding pronoun disfavor overt pronoun use (25.8%). The table also reveals that the presence of a preceding pronoun is much less common than

having no preceding pronoun ($Ns = 27,790$ & $60,211$, respectively).

	<i>N</i> VERBS	% OVERT PRONOUNS
Preceding pronoun present	27,790	44.2
Preceding pronoun absent	60,211	25.8

Table 4.22: Preceding Pronoun

To explore the nature of the relationship between pronoun use and Preceding Pronoun, a mixed effects logistic regression model was run. The model contained Preceding Pronoun as a fixed effect, verb as a random effect, and pronoun presence vs. absence as the response variable. The main effect model outputs for Preceding Pronoun are described in Table 4.23.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.46	0.02	-65.39	< 0.001***
Preceding pronoun present	0.81	0.02	49.55	< 0.001***

Table 4.23: Main effect of Preceding Pronoun; Model configuration: `glmer(Pronoun ~ Preceding Pronoun + (1 | Verb))`; Reference level: preceding pronoun absent

Model results indicate a significant effect of Preceding Pronoun ($\beta = 0.81, p < 0.001$) such that a pronoun is significantly more likely when a pronoun is present in the preceding site of pronominal variation. As with the other predictors, the effect is not particularly strong, with a marginal R^2 of 0.034. The conditional R^2 is 0.213. Nevertheless, this finding is in line with previous research on Preceding Pronoun (Cameron and Flores-Ferrán, 2004; Travis, 2007; Torres Cacoullos and Travis, 2011).

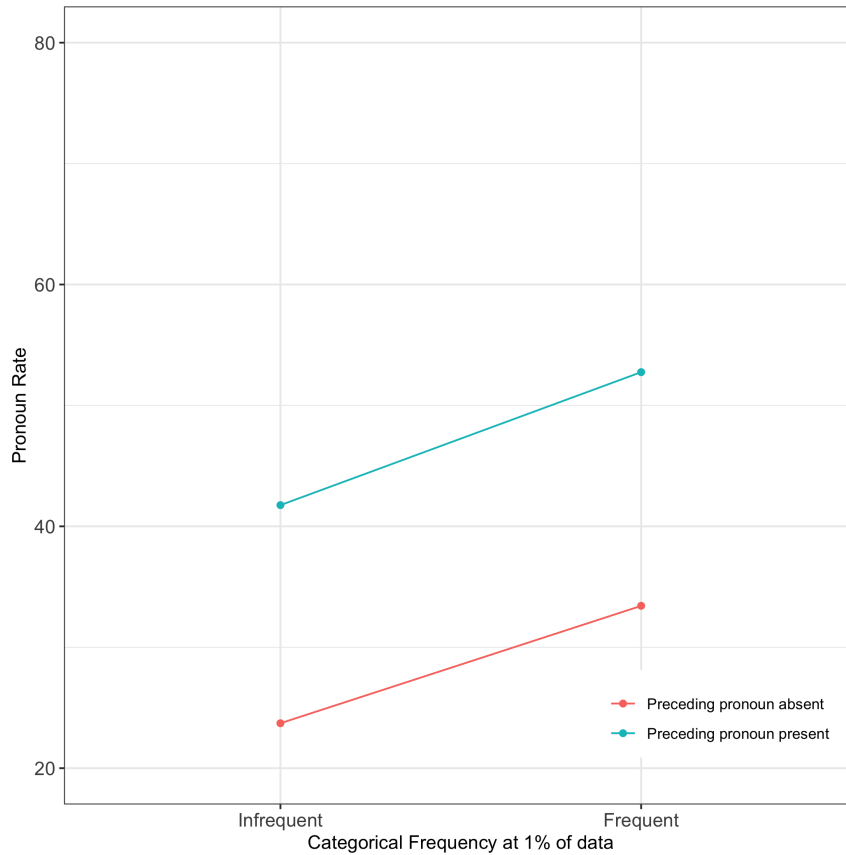


Figure 4.15: Preceding Pronoun: frequent vs. infrequent forms

Interaction with Discrete Frequency

This subsection explores the extent to which Preceding Pronoun interacts with Discrete Frequency. As with the interactions investigated in the replication of Erker and Guy (2012), this section presents the pronoun rates for each level of Preceding Pronoun (preceding pronoun present and preceding pronoun absent). These are visualized in Figure 4.15.

As the figure shows, pronoun rates for infrequent forms are lower than pronoun rates for frequent forms for all levels of Preceding Pronoun. For infrequent tokens with preceding pronoun present, the pronoun rate is 41.8%, but for infrequent forms with preceding pronoun absent the rate of pronoun use is 23.7%. For frequent forms,

the pronoun rate for preceding pronoun present is 52.8%, and the pronoun rate for preceding pronoun absent is 33.4%. Just like the predictors investigated in the replication, we see that Discrete Frequency amplifies the effect of Preceding Pronoun on pronoun use. Although small, there is an larger range in pronoun rates between frequent forms with preceding pronoun present and preceding pronoun absent than between infrequent forms of the same values ($Range_{freq} = 19.4$; $Range_{infreq} = 18.1$).

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.73	0.24	-3.00	0.003**
Preceding pronoun present	0.8	0.03	24.07	< 0.001***
Infrequent	-0.73	0.24	-3.02	0.003**
Preceding pronoun present*Infrequent	0.02	0.04	0.48	0.628

Table 4.24: Interaction between Preceding Pronoun and Discrete Frequency; Model configuration: `glmer(Pronoun ~ Preceding Pronoun*Discrete Frequency + (1 | Verb))`; Reference levels: preceding pronoun absent, frequent, preceding pronoun absent*frequent

Two mixed effects logistic regression models were run to determine the extent to which Preceding Pronoun interacts with Discrete Frequency. The first model includes Preceding Pronoun and Discrete Frequency as individual fixed effects and verb as a random effect. The second model includes Preceding Pronoun and Discrete Frequency as main effects and as an interaction, and it also includes verb as a random effect. Table 4.24 presents the model outputs for the interaction model. Model results indicate a significant effect for infrequent forms ($\beta = -0.73, p = 0.003$). Interestingly, there is no significant interaction between preceding pronoun present and infrequent forms ($\beta = 0.02, p = 0.628$).

The marginal R^2 for this model is identical to the regular model at 0.55, and the conditional R^2 for the interaction model is only 0.001 higher than the regular model at 0.228. This indicates that Preceding Pronoun and Discrete Frequency together

are only accounting for a small portion of the observed variation, and that does not change when they are included as an interaction. In other words, these results suggest that Preceding Pronoun is a maximally discursive factor that is not need amplification from Frequency to increase its impact. Further, an ANOVA comparing the regular model and the interaction model shows no significant difference between these models ($p = 0.63$), suggesting that there is not an interaction between Discrete Frequency and Preceding Pronoun. These results are interesting to account for, since the data in Figure 4-15 suggest an amplification effect that would be indicative of a significant interaction between priming and discrete frequency.

4.9.2 Morphological Regularity 2.0

Table 4.25 shows the pronoun rates and token counts for each level of morphological regularity 2.0. As the figure shows, there is minimal difference between the pronoun rates in each of the three categories. Nevertheless, a clear trend is observed: regular verbs have the highest overt pronoun rates (32.2%), followed by semi-irregular verbs (31.3%), and then irregular verbs (30.7%). While these levels don't differ drastically in their pronoun rates, they do present large differences in their frequencies of occurrence. Regular forms are most frequent ($N = 45,675$), then irregular forms ($N = 23,368$), and finally semi-irregular forms ($N = 18,958$).

	<i>N</i> VERBS	% OVERT PRONOUNS
Regular	45,675	32.2
Semi-irregular	18,958	31.3
Irregular	23,368	30.7

Table 4.25: Morphological Regularity 2.0

A mixed effects regression model was run that included morphological regularity as a fixed effect, verb as a random effect, and pronoun as the response variable. Model

results (presented in Table 4.26) indicate a significant effect of semi-irregular forms ($\beta = 0.24, p = 0.003$), such that they are more likely to occur with a pronoun than the reference level Irregular.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.24	0.05	-22.81	< 0.001***
Regular	0.03	0.06	0.56	0.57
Semi-irregular	0.24	0.08	2.97	0.003**

Table 4.26: Main effect of Morphological Regularity 2.0; Model configuration: `glmer(Pronoun ~ Morphological Regularity 2.0 + (1 | Verb))`; Reference level: irregular

Overall, this variable presents a very weak effect on pronoun production ($R_M^2 = 0.002$), while the conditional R^2 is 0.196. It does seem that this operationalization provides a clearer understanding of morphological regularity as it pertains to pronoun production, since there is a statistically significant result here, while there was not one for binary morphological regularity (see Section 4.3).

Interaction with Discrete Frequency

Now we shall investigate the extent to which morphological regularity 2.0 interacts with Discrete Frequency. Figure 4.16 presents the pronoun rates by Discrete Frequency for morphological regularity 2.0. As the figure shows, there is an increase in pronoun rates for regular verb forms when changing from infrequent to frequent, but there is no such increase for semi-irregular or irregular forms¹¹. This is somewhat expected, since the replication showed a similar result for binary Morphological Regularity (see Figure 4.7).

¹¹Neil Myler (pers. comm.) observes that, although there is widespread understanding that speakers must memorize specific semi-irregular and irregular verb forms, the results presented here

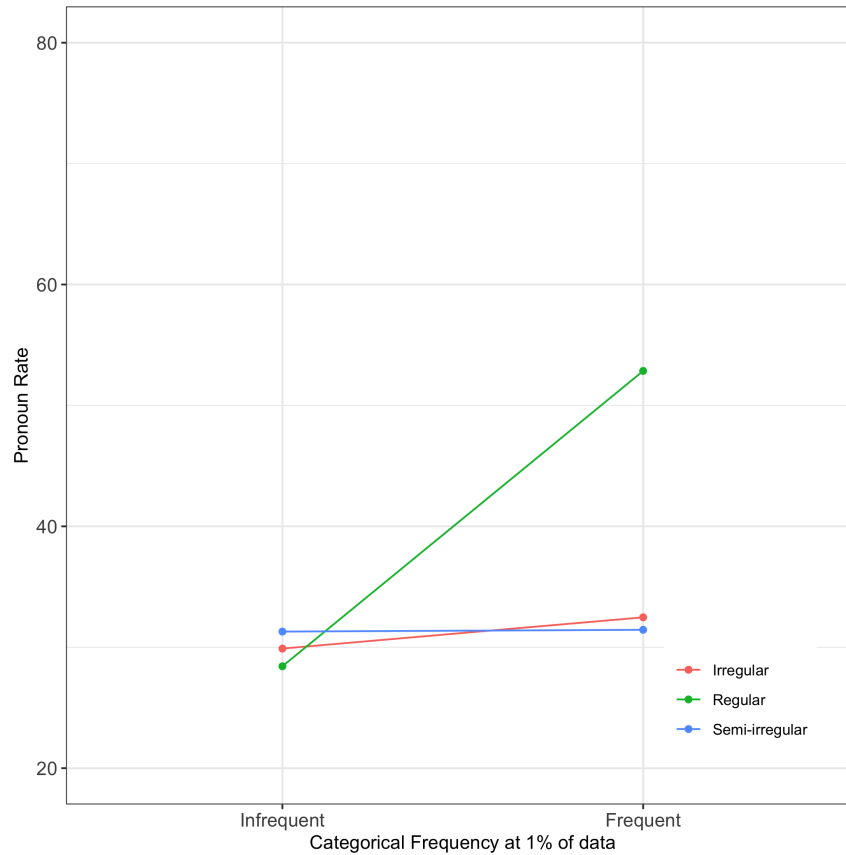


Figure 4.16: Morphological Regularity 2.0: Frequent vs. infrequent forms

As with previous linguistic predictors, two mixed effects regression models were run to evaluate the potential interaction between morphological regularity 2.0 and Discrete Frequency. The first model included pronoun presence as the response variable, morphological regularity 2.0 and Discrete Frequency as main effects, and verb as a random effect. The second model had nearly an identical construction as the first, but included morphological regularity 2.0 and Discrete Frequency as an interaction. Model outputs for the interaction model are presented in Table 4.27.

Interaction model results show a significant effect at the Regular*Infrequent level ($\beta = -0.806, p = 0.049$). The other interaction level and the Infrequent main effect suggest that the pronominal tendencies of those forms are seemingly not memorized.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.754	0.320	-2.356	0.018
Regular	0.798	0.406	1.962	0.05*
Semi-irregular	0.178	0.587	0.304	0.761
Infrequent	-0.449	0.324	-1.388	0.165
Regular*Infrequent	-0.806	0.409	-1.969	0.049*
Semi-irregular*Infrequent	0.018	0.592	0.030	0.976

Table 4.27: Interaction between Morphological Regularity 2.0 and Discrete Frequency; Model configuration: `glmer(Pronoun ~ Morphological Regularity 2.0*Discrete Frequency + (1 | Verb))`; Reference levels: irregular, frequent, irregular*frequent

level are not significant. The marginal R^2 for the interaction model is 0.031, and the conditional R^2 for the interaction model is 0.216. An ANOVA comparing the interaction model to the regular model reveals that these two models are not significantly different from each other (Regular model AIC = 102,404; Interaction model AIC = 102,405; $p = 0.202$). This indicates the absence of an interaction effect between morphological regularity 2.0 and Discrete Frequency.

4.9.3 Semantic Category 2.0

We now turn to the four-way coded semantic category variable. As shown in Table 4.28, estimative verbs are most likely to occur with a pronoun (51.1%). Mental activity verbs correspond to the second highest overt pronoun rate (37.6%). Stative verbs have the third highest pronoun rate (32.4%), followed by external activity verbs (28.6%). Interestingly, we see an inverse relationship between pronoun rate and frequency of occurrence with this variable: the lowest pronoun rates correspond to the most-frequent semantic category. External activity verbs make up the largest

category, with 48,026 observations, followed by Stative verbs ($N = 23,380$), then Mental activity verbs ($N = 14,664$), and finally Estimative verbs ($N = 1,931$). It is important to note that nearly half of the Estimative verbs (47.7%) are essentially just *creo* ‘I believe’ ($N_{creo} = 917$). The second most frequent form in the estimative group is *pienso* ‘I think’, with 80 observations.

	N VERBS	% OVERT PRONOUNS
Estimative	1,931	51.1
Mental Activity	14,664	37.6
Stative	23,380	32.4
External Activity	48,026	28.6

Table 4.28: Semantic Category 2.0

A mixed effects model including semantic category as the sole fixed effect was run to determine the statistical impact of this variable on pronoun production. As with all previous models, verb was included as a random effect. The model outputs are displayed in Table 4.29.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.65	0.06	-10.32	< 0.001***
External Activity	-0.57	0.06	-9.02	< 0.001***
Mental Activity	-0.49	0.06	-8.41	< 0.001***
Stative	-0.46	0.06	-7.10	< 0.001***

Table 4.29: Main effect of Semantic Category 2.0; Model configuration: `glmer(Pronoun ~ Semantic Category 2.0 + (1 | Verb))`; Reference level: estimative

Model results indicate highly significant effects for all levels ($ps < 0.001$). Specifically, external activity verbs, mental activity verbs, and stative verbs are all signifi-

cantly less likely to occur with a pronoun compared to the reference level (Estimative verbs). This suggests that semantic verb category does impact pronoun production. Nonetheless, the effect is weak, with a marginal R^2 of only 0.002, and a conditional R^2 of 0.194.

Interaction with Discrete Frequency

Figure 4.17 presents the pronoun rates for Semantic Category 2.0 by Discrete Frequency. The figure shows higher pronoun rates for all levels of semantic category 2.0 for frequent forms compared to infrequent forms. Although slight, the figure does show an amplification effect of Discrete Frequency: the range in rates for infrequent forms ($Range_{infrequent} = 17.4$) is lower than the range in rates for frequent forms ($Range_{frequent} = 21.2$). Estimative verbs have the highest pronoun rates for infrequent and frequent forms (45.7% and 56.7%, respectively). External activity verbs correspond to the lowest pronoun rates for infrequent and frequent forms (28.3% and 35.5%, respectively).

To determine the extent to which an interaction is occurring between semantic category 2.0 and Discrete Frequency, two mixed effects logistic regression models were run. The first model contained semantic category 2.0 and Discrete Frequency as separate main effects and verb as a random effect. The second model included semantic category 2.0 and Discrete Frequency as main effects and as an interaction, and it also included verb as a random effect. Results from the interaction model are displayed in Table 4.30.

The results of the interaction model show that all three interaction levels are statistically significant ($ps < 0.001$). The main effect of discrete frequency is not significant ($p = 0.866$). The Marginal R^2 for this model is 0.024, suggesting that the interaction accounts for a small portion of the observed variation in the data. Nevertheless, an ANOVA comparing the regular model and the interaction model

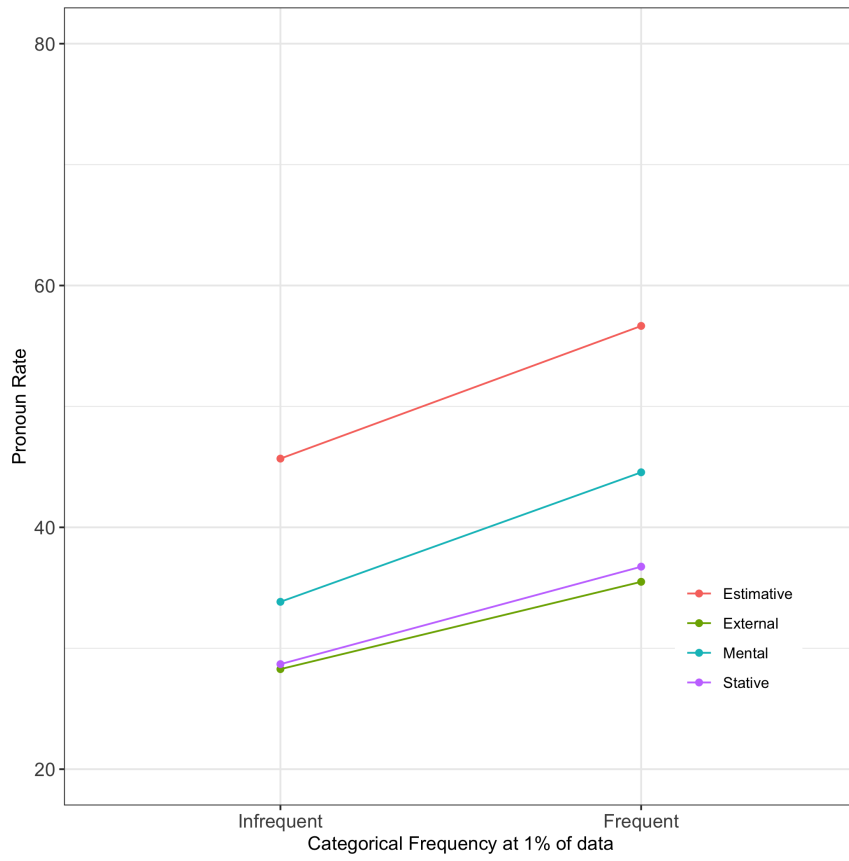


Figure 4-17: Semantic Category 2.0: Infrequent and frequent forms

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.352	0.240	-1.466	0.143
External Activity	0.053	0.127	0.421	0.674
Mental Activity	-0.271	0.078	-3.491	< 0.001***
Stative	-0.070	0.098	-0.716	0.474
Infrequent	-0.042	0.251	-0.169	0.866
External Activity*Infrequent	-0.886	0.150	-5.915	< 0.001***
Mental Activity*Infrequent	-0.461	0.114	-4.056	< 0.001***
Stative*Infrequent	-0.663	0.130	-5.098	< 0.001***

Table 4.30: Interaction between Semantic Category 2.0 and Discrete Frequency; Model configuration: `glmer(Pronoun ~ Semantic Category 2.0*Discrete Frequency + (1 | Verb))`; Reference levels: estimative, frequent, estimative*frequent

shows a significant difference between them ($p < 0.001$), and the interaction model has an AIC that is 27 points lower than the regular model ($AICs = 102,300$ and $102,327$, respectively). These findings indicate that semantic category 2.0 does interact with Discrete Frequency. These results are not particularly surprising, since there was a significant interaction between Semantic Category and Discrete Frequency in the replication (see Section 4.6), and Semantic Category 2.0 simply splits one of the three levels into two. The marginal R^2 for Semantic Category 2.0 is higher than the marginal R^2 of Semantic Category (0.024 and 0.023, respectively), confirming that this four level operationalization is more explanatory than that of Erker and Guy (2012).

4.10 Summary of All Linguistic Predictors

Table 4.31 presents a summary of the results for the linguistic predictors from the replication and the three additional predictors presented above. Just like Table 4.20, this table includes the main effects, interactions with frequency, and evidence of amplification effects. Interestingly, Preceding Pronoun and Morphological Regularity 2.0 show no statistically significant interaction with Discrete frequency even though they do show amplification effects via higher ranges for frequent than infrequent forms.

CONDITIONING FACTOR	MAIN EFFECT	INTERACTION W/FREQUENCY	EVIDENCE OF AMPLIFICATION EFFECT
Morph Regularity	NO $p = 0.112$	YES $p = 0.035$	YES Range diff. = 18.7 (larger in freq. forms)
Person/Number	YES $p < 0.05$	YES $p = 0.003$	YES Range diff. = 28.6
TMA	YES $p < 0.05$	NO $p = 0.12$	NO Range diff. = -4.1
Semantic content	YES $p < 0.001$	YES $p < 0.001$	YES Range diff. = 5.2
Switch reference	YES $p < 0.001$	YES $p < 0.001$	YES Range diff. = 5.5
Preceding Pro	YES $p < 0.001$	NO $p = 0.63$	YES Range diff. = 1.3
Morph Regularity 2.0	YES $p = 0.006$	NO $p = 0.202$	YES Range diff. = 20.0
Semantic Category 2.0	YES $p < 0.001$	YES $p < 0.001$	YES Range diff. = 3.8

Table 4.31: Summary of main effects, interactions with frequency, and evidence of amplification effects for all linguistic constraints

In the next chapter, I expand upon the results from the replication presented here through a detailed investigation into the contextual frequency metrics. The main goal in this expansion is to obtain a better understanding of the fine-grained detail that speakers may attune to when storing information about verb forms and their pronoun rates. The expansion in Chapter 5 also includes a section that joins predictor variables from the replication with contextual frequency metrics. Combining multiple linguistic constraints with contextual frequency metrics allows us to better understand how these predictors work together to impact pronoun production.

Chapter 5

Expanding Erker & Guy (2012): Analysis of Contextual Frequency Metrics

The replication results provide encouraging support for the role of lexical frequency as an amplifier of the effects of linguistic constraints in morphosyntactic variation. However, as discussed in previous chapters, the frequency metrics investigated in the replication could be enriched by including information about the context in which the verbs occurred, a key insight of UBG. For this reason, this chapter presents the expanded analysis of frequency. Results from this portion of the study reveal that the contextual frequency metrics are generally better predictors of subject pronoun presence than the overall-frequency measures studied in Erker and Guy (2012). Nevertheless, this investigation of contextual frequency metrics also indicates that overall frequency matters, too. The data suggest that finite verb forms must reach a certain overall frequency threshold for sensitivity to contextual information to appear. The increased explanatory power of contextual frequency metrics, coupled with more in-depth investigation of these contextual frequency metrics support the UBG notion that context has an impact on usage patterns, which in turn impact mental representations. In other words, results indicate that language users acquire information about the contextual tendencies of finite forms through usage (measured via contextual frequencies), and, when these forms are sufficiently frequent, these contextual frequencies modulate the pronominal tendencies for that form.

5.1 General Description of Contextual Frequency

As a reminder, the novel contextual frequency metrics presented in the present study consist of the frequency at which each verb appears in each of four Switch Reference/Preceding Pronoun contexts: ‘different referent/preceding pronoun absent’, ‘different referent/preceding pronoun present’, ‘same referent/preceding pronoun present’, and ‘same referent/preceding pronoun absent’. In addition to these four frequency metrics, I also present results for the Favoring Context Ratio (FCR) and Disfavoring Context Ratio (DCR) (see Sections 5.5 and 5.6), which are the ratios at which each verb appears in the pronoun-favoring and pronoun-disfavoring contexts relative to the overall frequency of each verb, respectively. First, we will look at the general trends of the four contextual frequencies in the dataset.

	<i>N</i> VERBS	% OF CORPUS	% OVERT PRONOUNS
Different referent/Preceding pronoun present	14,965	17.0	46.3
Same referent/Preceding pronoun present	12,825	14.6	41.7
Different referent/Preceding pronoun absent	32,054	36.4	33.6
Same referent/Preceding pronoun absent	28,157	32.0	16.9

Table 5.1: Contextual Frequency metrics

Table 5.1 presents the total number of verbs that appear in each of the contexts, the percent at which each context appears in the corpus, as well as their respective pronoun rates. Generally speaking, the results in Table 5.1 align with expectations: what was purported to be the pronoun-favoring context (‘different/present’) has the highest pronoun rate, what was purported to be the pronoun-disfavoring context (‘same/absent’) has the lowest pronoun rate, and the mixed contexts have mid-range pronoun rates. The table shows that each of these contexts is not symmetrically represented in the dataset. Verbs are more likely to appear in some of these contexts than others, which is not surprising in some respects, but was not entirely predictable

in other ways, at least in terms of overall pronoun rates. The number of instances in which a verb appears in either Preceding Pronoun context is generally straightforward to predict since it corresponds to the overall pronoun rate in the data. An overall pronoun rate of 31.6% in the current dataset signifies that 31.6% of tokens appear with a value of ‘present’ for Preceding Pronoun. This also indicates that 68.4% of tokens appear with a value of ‘absent’ for Preceding Pronoun. However, what is not predictable in a similar way is how often referents change from verb to verb, i.e. the frequency at which verbs occur with a value ‘different’ or ‘same’ for Switch Referent. That information is obtainable from the different rates of the four contexts presented in Table 5.1. The different/absent context is the most frequently occurring ($N = 32,054$, or 36.4% of all tokens). The pronoun-disfavoring context (same/absent) is the second most frequent ($N = 28,157$, or 32% of the corpus). The pronoun-favoring context (different/present) is the third most frequent context in the corpus ($N = 14,965$), representing only 17% of the dataset. Finally, the same/present context appears at the lowest frequency ($N = 12,825$, i.e. 14.6% of all tokens). Interestingly, the pronoun-favoring and pronoun-disfavoring contexts have pronoun rates that are equidistant from the overall pronoun rate in the corpus. That is to say that in the OZC-BSC corpus, which has an overall pronoun rate of 31.6%, a strong favoring or disfavoring context results in a pronoun rate that is increased or decreased 14.7% relative to the average rate.

At the most basic level, the results in Table 5.1 provide clear evidence that the potentially pronoun-favoring and pronoun-disfavoring contexts are, in fact, pronoun-favoring and pronoun-disfavoring. Further, the results in Table 5.1 act as an indirect ranking of the effect size of Switch Reference and Preceding Pronoun, since the two mixed contexts have quite different pronoun rates. The predictions in Chapter 3.5 were such that if the mixed Switch Reference and Preceding Pronoun contexts equally

impacted pronoun use, their pronoun rates would be identical. However, we see now that that prediction is not borne out. As previously mentioned, the overall pronoun rate in the corpus is 31.6%. The two mixed context pronoun rates are 41.7% ('same referent/preceding pronoun present') and 33.6% ('different referent/preceding pronoun absent'). The pronoun rate for the same referent/preceding pronoun present mixed context is much higher than the overall pronoun rate, while the pronoun rate for the different referent/preceding pronoun absent mixed context is much closer to the overall rate. This indicates that Preceding Pronoun has a stronger effect on pronoun use than Switch Reference, since the mixed context with the higher pronoun rate contains the Preceding Pronoun level that favors pronoun use (preceding pronoun present).

Now that we have a sense of the overall distribution of the observations and pronoun rates across these metrics, let us now ask how they relate to questions of frequency. What might incorporating these contextual properties into frequency metrics tell us about pronoun use in general? Might it be the case that two forms with very different pronoun rates are differentially frequent in their occurrence across these four contexts? In other words, to what extent do the most-frequent contexts of verb forms impact their overall pronominal tendencies? These are the driving questions that will be considered as the study transitions to a more detailed investigation of the six contextual frequency metrics. The objective for this investigation is to determine whether the pronoun rates of specific verbs are predictable on the basis of how often, proportionally speaking, they occur in specific contexts known to (dis)favor pronoun use. If contextual frequency metrics do, in fact, impact pronoun rates, this would support usage-based accounts of the detailed nature of the grammar.

The next section examines the relationship between Lexical and Contextual Frequency in order to determine the extent to which Lexical Frequency and its effects

impact Contextual Frequency. Section 5.3 presents an analysis of the four Log Contextual Frequency Metrics that outlines the problems with investigating Log Frequencies in this manner. Then, in Section 5.4, a non-linear investigation of contextual frequency is explored through the pronoun-disfavoring log frequency. Sections 5.5 and 5.6 present the main thrust of Contextual Frequency analysis, with the investigations of FCR and DCR.

5.2 Relationship Between Lexical and Contextual Frequency

This section aims to better understand the relationship between Lexical Frequency and the Contextual Frequency metrics defined in this study. As previously mentioned, contextual frequencies were calculated by taking the log-transformation of the number of times each finite form appeared in the four Switch Reference/Preceding Pronoun contexts. For example, the log frequency Different referent/Preceding pronoun present consists of the log transformed frequency at which each finite form appeared with a switch in referent and with a preceding pronoun marked as ‘present’. Since Contextual Frequency Metrics are simply the log-transformed frequencies at which finite forms appear in each of the four Switch Reference/Preceding Pronoun contexts, it is likely that the overall frequency effect observed in the previous chapter (despite its inconsistency) will have an impact on the contextual frequency effects. For instance, a verb that appears 100 times will have some combination of occurrences across the four contexts that will, when summed, add up to 100. While there is no clear way to predict how many times a form will appear in each of these contexts, it is sure that the number of occurrences in each context will be much higher than the number of occurrences for a verb with an overall frequency of 10. The higher overall frequency of the verb that appears 100 times guarantees that the counts of its contextual frequencies will be higher than those of a verb that appears only 10 times. To

illustrate this quantitatively, a correlation matrix that compares Log Frequency and the four Log Contextual Frequencies was generated. The results from the correlation test are outlined in Table 5.2.

	Log Freq.	Different/Present	Same/Absent	Same/Present	Different/Absent
Log Freq.	1.00	0.97***	0.97***	0.94***	0.98***
Different/Present		1.00	0.93***	0.92***	0.98***
Same/Absent			1.00	0.95***	0.95***
Same/Present				1.00	0.92***
Different/Absent					1.00

$p < 0.001$ '***'

Table 5.2: Correlation matrix for Log Frequency and four Contextual Frequency Metrics

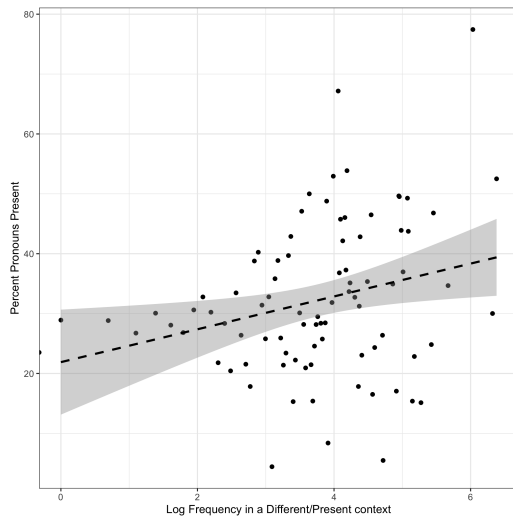
As these results show, all five log frequencies (Log Frequency and the four Log Contextual Frequencies) are strongly correlated with each other. The two strongest correlations (at 0.98) are between Log Frequency and Log Frequency Different/Absent, and Log Frequency Different/Present and Log Frequency Different/Absent. The two lowest correlations, which are still very strong at 0.92, are between Log Frequency Different/Present and Log Frequency Same/Present, and Log Frequency Same/Present and Log Frequency Different/Absent. These strong correlations are unsurprising since the log frequency is essentially the log transformed sum of the raw contextual frequencies for a given verb form. Each instance of a finite form appearing in a given context also corresponds to an instance of use, such that the contextual frequency increases as a forms overall frequency increases. This correlation matrix highlights the challenge in investigating contextual frequency effects. It is not feasible to simply replace a lexical frequency metric with a contextual frequency because contextual frequency (operationalized as a kind of count) is a function of overall verb frequency. The next section, which outline the results for the four contextual log frequencies, underscores the inadequacies that arise when investigating (non-ratio) contextual frequency

measures given the way they've been defined here. Section 5.4 explores this further by looking at another way of binning tokens. Together, these sections underscore the complexity that exists in examining contextual frequency due to its relationship with overall frequency, and point to the ratio-based frequencies as better metrics to disentangle this relationship.

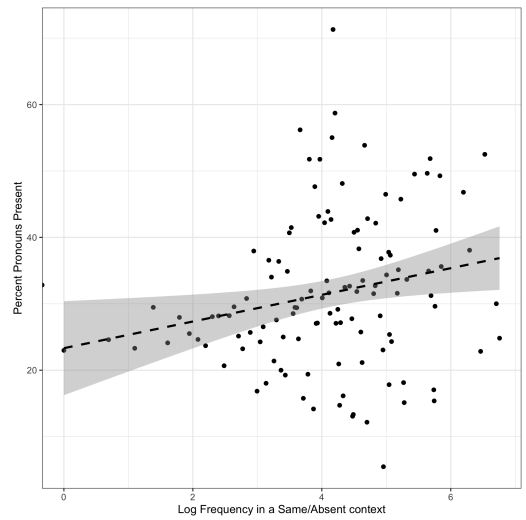
5.3 Log Contextual Frequencies

Figure 5.1 plots the pronoun rates based on the log frequency of the verb appearing in all four combinations of factor values for the variables Switch Reference and Preceding Pronoun. Figure 5.1a plots the log frequency of the verb appearing in the combination of factor values for the variables Switch Reference and Preceding Pronoun that are expected to favor pronoun use (the verb in question (a) constitutes a change in referent and (b) follows a site of pronominal variation in which a subject pronoun was present). Figure 5.1b plots the log frequency of the verb appearing in the context that is pronoun-disfavoring with respect to Switch Reference and Preceding Pronoun, such that a given verb constitutes no change in referent and follows a site of pronominal variation in which a subject pronoun was not present. Figures 5.1c and 5.1d plot the pronoun rates for the log frequencies of the verbs in question appearing in the two mixed contexts: without a switch in referent and following a site of pronominal variation in which a subject pronoun was present; and with a switch in referent and following a site of pronominal variation in which a subject pronoun was not present.

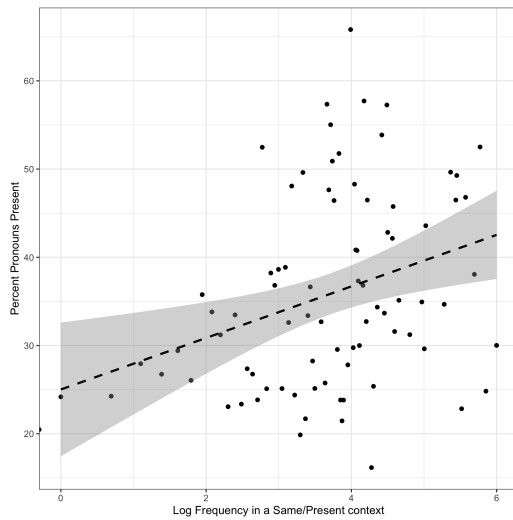
The results plotted in Figure 5.1 show that pronoun rates increase as the contextual frequencies increase. However, similar to the Log Frequency figure in Chapter 4, the figures here show quite a bit of variation in pronoun rates that is concentrated on the right side of each plot indicating increased variability for higher frequency forms. Given that the previous section revealed that the log contextual frequencies are near



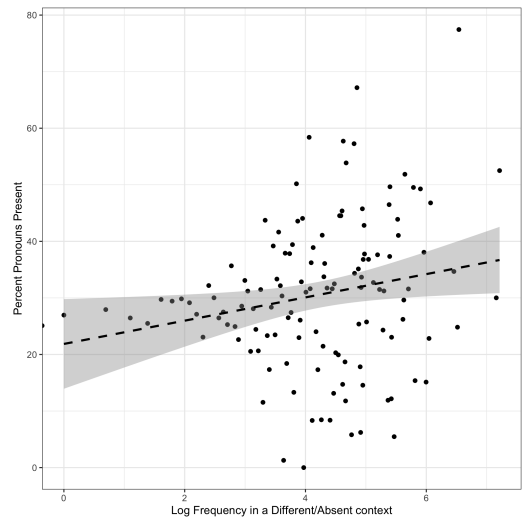
(a) Log Pronoun Favoring



(b) Log Pronoun Disfavoring



(c) Log Mixed Context 1



(d) Log Mixed Context 2

Figure 5.1: Log Contextual Frequencies and Percent Pronouns Present

	PRONOUN RATE
Log Freq. Different/Present	0.25*
Log Freq. Same/Absent	0.22*
Log Freq. Same/Present	0.31**
Log Freq. Different/Absent	0.20*

$p < 0.05$ ‘*’; $p < 0.01$ ‘**’

Table 5.3: Correlation matrix of pronoun rates for the four Contextual Frequency Metrics

identical to overall frequency, the results presented here are unsurprising. This figure and the correlation matrix presented in Table 5.3 show substantial evidence that a linear fit to these data, although statistically significant (and suggestive of some kind of overall frequency effect), cannot adequately illuminate the nature of any of the frequency metrics, contextual or otherwise, as they impacts pronoun use in Spanish. As the figure shows, a linear fit of the data is very poorly predictive of the behavior of some of the higher frequency forms that have really low pronoun rates and others that have much higher pronoun rates. Additionally, the correlation matrix in Table 5.3 shows that the correlations between each contextual frequency metric and their respective pronoun rates are weak.

In the following section, I explore this further through an investigation of pronoun disfavoring frequency that includes a different way of binning tokens. Binning tokens based on overall frequency allows for a deeper dive into potential contextual frequency effects that are non-linear in nature.

5.4 Exploring Contextual Frequency Non-Linearly

This section presents an exploratory analysis of contextual frequency. The goal here is to use a different method for binning tokens in an effort to make sense of the

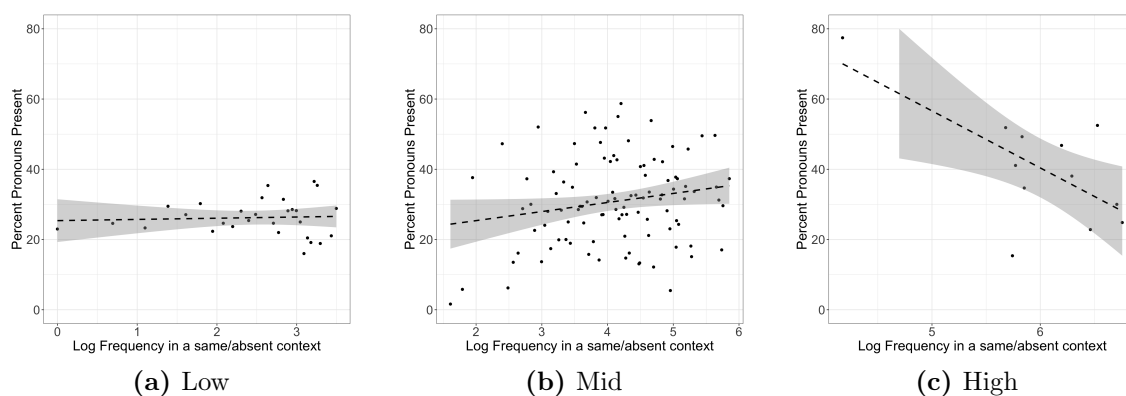


Figure 5-2: Percent pronouns present by Log Frequency of appearing in a pronoun disfavoring context for three different lexical frequency groups

non-linear relationship between contextual frequency and pronoun use. Figure 5-2 compares the pronoun rates for three different groups of verb forms: (a) verb forms that appear less than 55 times in the corpus (low frequency), (b) forms that appear between 55 and 879 times and will be referred to as *mid-range frequency*, and (c) the 12 high frequency forms from the replication that constitute at least 1% of the corpus (appearing 880 times) and were labelled as “frequent” for Discrete Frequency¹. Dividing the data into these groups provides increased clarity.

The figure shows that the frequency of occurrence in a disfavoring context impacts the three lexical frequency groups in different ways. The low frequency group seems to show nearly no effect of disfavoring context frequency (5-2a). As the log frequency of appearing in a disfavoring context increases, pronoun rates seemingly remain the same for low-frequency forms. For the mid-frequency forms, Figure 5-2b suggests a positive relationship between pronoun disfavoring frequency and pronoun rate: the more a mid-frequency verb form appears in a pronoun disfavoring context, the higher the pronoun rate of that form. Finally, for the high frequency group, (5-2c), the

¹This three-way distinction should not be confused with the binary lexical frequency metrics DISCRETE FREQUENCY, which divided verb forms as “frequent” and “infrequent” based on whether they constituted at least 1% of the corpus.

pronoun-disfavoring contextual frequency does associate with pronoun rates in the expected direction. As log frequency of appearing in the pronoun disfavoring context increases, pronoun rates for frequent forms decrease.

Three correlation statistics were run to investigate the relationship between pronoun-disfavoring frequency and pronoun rate for each of the groups depicted in Figure 5·2. The correlation results for the frequent forms show a significant negative correlation between the frequency of appearing in a pronoun-disfavoring context and pronoun rate ($r(10) = -0.67; p = 0.017$). The correlation results for the mid-frequency forms show a positive correlation between frequency of occurring in a pronoun-disfavoring context and pronoun rate, but it is only near significance ($r(100) = 0.19; p = 0.053$). Finally, for the low frequency forms, the correlation results indicate a very weak positive correlation between pronoun-disfavoring frequency and pronoun rate, but it is not statistically significant ($r(27) = 0.06; p = 0.77$).

One possible explanation for the results in Figure 5·2 is that the different overall frequencies impact the extent to which lexical *and* contextual frequencies manifest. For high frequency forms, their exemplar clouds contain enough exemplars that they have accrued a sensitivity to context, which is why we see a robust effect of the disfavoring context frequency in the expected direction. The mid-frequency group (shown in 5·2b) does not show the same contextual frequency effect, since the linear fit indicates that pronoun rate increases as the frequency of occurrence in a disfavoring context increases. However, this increase in pronoun rate is consistent with the overall lexical frequency effects described in Section 4.2 (an increase in frequency corresponds to a small increase in pronoun use). As with Figure 4·3, there is substantial variation in pronoun rates, and most points plotted in Figure 5·2b are outside of the linear fit. The low frequency group presents pronoun rates that remain nearly unchanged as pronoun-disfavoring frequency increases along the x-axis. Overall, this figure suggests

that lexical *and* contextual frequencies impact the pronominal tendencies of verbs. Very infrequent forms have such small exemplar clouds that they are not sensitive to contextual frequencies and present minimal variation in their pronoun rates. As overall frequency increases, mental representations acquire more detail surrounding the usage patterns of verb forms. At a certain overall frequency, highly frequent forms have robust enough exemplar clouds and maintain sensitivity to contextual frequency effects.

The sometimes inconsistent, and evidently non-linear, relationship between contextual frequency, overall lexical frequency, and pronoun use are the exact difficulties that the FCR and DCR attempt to avoid. With this in mind, we now turn to an investigation of FCR (Section 5.5) and DCR (Section 5.6).

5.5 Favoring Context Ratio

The ratio-based frequency metrics (FCR and DCR) bypass much of the complexities that lexical frequency effects have presented above because they are proportions that account for differences in overall frequencies. As a reminder, the FCR is the ratio at which a verb appears in the pronoun favoring context (with a different referent and preceding pronoun present) relative to the overall frequency of the verb. As described in Section 3.4, FCR is calculated by dividing the frequency at which the verb appears in the Different/Present context by the total frequency of the verb. This contextual frequency metric better accounts for the discrepancies between verb frequencies within the corpus, since in a sense it normalizes all frequencies to a single scale. It allows for a comparison of forms with different overall frequencies in terms of their distribution across four different pronominally relevant contexts. For example, *sé* ‘I know’ and *hacen* ‘they do’ occur at different overall frequencies (3,062 and 223, respectively), but they have the same FCR, 0.18.

Figure 5-3 plots the pronoun rates for each FCR. Interestingly, the trend line shows pronoun rates decreasing as FCR increases. This is contrary to what we would predict. We would predict that a higher FCR would correspond to higher overt pronoun rates. The confidence interval (in grey) on the figure is quite wide as the FCR increases, suggesting that there is less confidence in the slope of the trend line in this area. This is due to the relatively small number of verbs that have FCRs above 0.5.

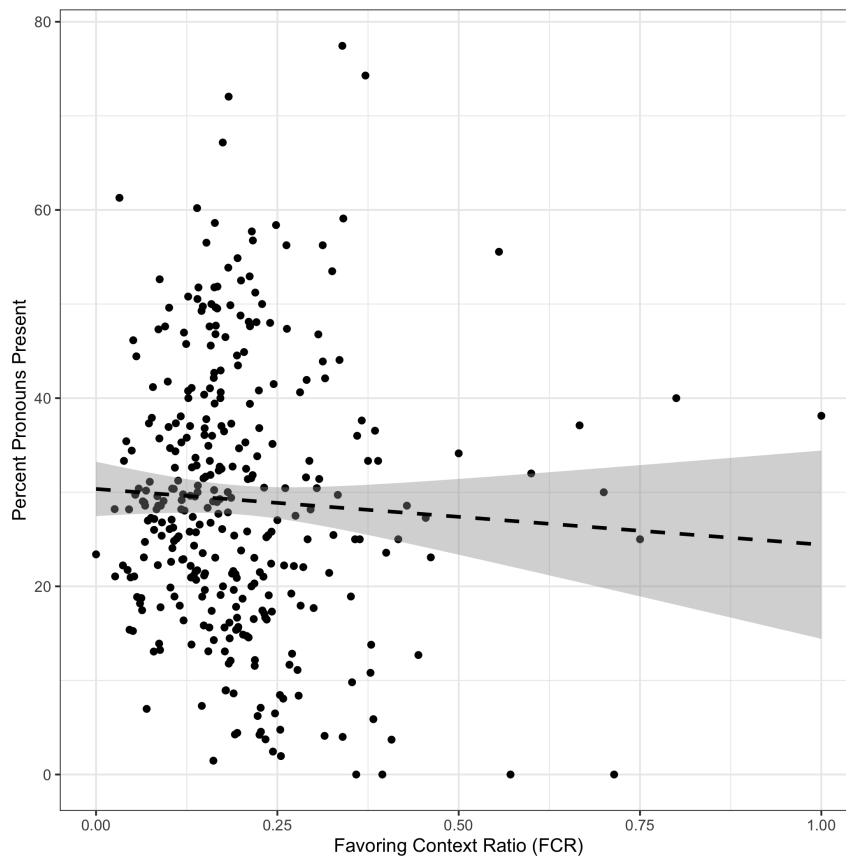


Figure 5-3: Favoring Context Ratio and percent pronouns present

In an effort to increase visual clarity, Figure 5-4 plots FCR values that are rounded to the second decimal point². Rounding inherently creates a simpler figure, since it

²It is unclear what the methodological implications are for adjusting the FCR in this manner. Does a difference in FCR of 0.00001 matter to a language user? This question, while interesting and

bins more FCRs together. However, regardless of whether FCR is presented rounded or unrounded, the variable still behaves in a way that is opposite to our prediction: higher FCR appears to correspond to *lower* pronoun rate.

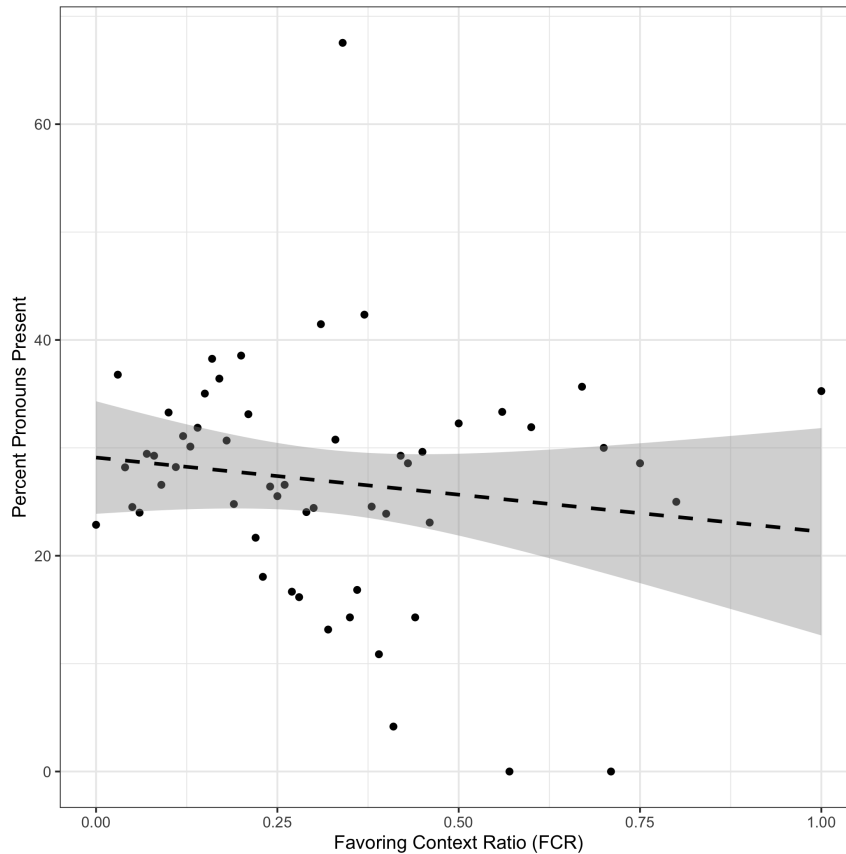


Figure 5·4: Favoring Context Ratio (rounded to two decimal points) and percent pronouns present

As with the previous contextual frequency metrics, a Pearson’s correlation statistic was run to investigate the relationship between unrounded FCR and the percent of pronouns present. Correlation results indicate that there is a negative correlation, though not statistically significant ($r(333) = -0.052$, $p = 0.3453$). This result is consistent with the data as visualized in Figures 5·3 and 5·4, but it is nevertheless surprising.

important, is outside of the scope of the current dissertation project.

As it currently stands, Figure 5.3 is somewhat challenging to interpret within a UBG framework. The linear fit of the figure amounts to a strong challenge of contextually-based frequency claims. Given the results in the previous section, which reveal contextual frequency effects exclusively for forms with high overall frequency, one possible explanation for this is that FCR alone is not sufficient in capturing the contextual frequency effects on verb forms. This is particularly insightful when we consider the fact that forms with identical FCRs may have very different overall frequencies. A low-frequency verb form with a high FCR is still potentially infrequent enough that the mental representation for that form (which informs language use) is not impacted by its FCR. Since each linguistic experience is stored in the exemplar cloud, the mental representation of a low-frequency form, which has fewer exemplars in its exemplar cloud, could therefore have insufficient information surrounding the contextual preferences of that form. This speculation leads to a testable prediction: if overall frequency impacts mental representations, the FCR *should* show the expected relationship for at least moderate to highly frequent forms. In other words, FCR is hypothesized to only impact the pronominal tendencies of a verb form if that form has already occurred at a rate that surpasses a frequency threshold, making it higher in frequency and therefore better represented.

To test this prediction, the dataset was divided into four subgroups: two based on FCR (one low FCR group between 0.1 and 0.2 and one high FCR group between 0.4 and 0.5) and two based on log frequency (one low log frequency group and one high log frequency group). Binning finite forms by log frequency should show different effects of FCR for the high and low log frequency groups. As mentioned above, one idea informed by usage-based thinking is that only the high log frequency group will present the expected positive correlation between FCR and pronoun rate, since the finite forms in this group should correspond to large-enough exemplar clouds.

In contrast, correlation results for the low log frequency group should *not* show a correlation between FCR and pronoun rate, since UBG would predict that the finite forms in this group are not frequent enough to develop sensitivities to contextual information such as FCR.

Correlation tests were run for each of the four groups: correlations between pronoun rate and log frequency were run for the FCR groups and correlations between pronoun rate and FCR were run for the log frequency groups. Results of these correlation tests are shown in Table 5.4.

GROUPED BY	GROUP	CORRELATION
Log Frequency	High Log Freq. (between 7.0 and 8.0)	0.88, $p < 0.001$
	Low Log Freq. (between 2.0 and 3.0)	0.04, $p = 0.001$
FCR	High FCR (between 0.4 and 0.5)	-0.15, $p < 0.001$
	Low FCR (between 0.1 and 0.2)	0.26, $p < 0.001$

Table 5.4: Summary of correlations between FCR and pronoun rate for low log frequency and high log frequency groups and correlations between log frequency and pronoun rate for low FCR and high FCR groups.

The correlation results for the low-FCR group show a significant positive correlation between log frequency and pronoun rate ($r(52,977) = 0.26, p < 0.001$). In other words, as log frequency increases, pronoun rate increases for forms with low FCR. This is an overall frequency effect that we would expect, given that this data show some evidence of a small, although weak, positive effect of overall frequency. However, when we look at the correlation between log frequency and pronoun rate for the high-FCR group, we see very different results. Correlation results indicate a negative, albeit weak, correlation between log frequency and pronoun rates for high-FCR forms ($r(1,340) = -0.15, p < 0.001$). Not only is this in the opposite direction of the trend that we would expect for an overall effect of frequency, but it is also against

the expected effect of FCR – which should correspond to higher pronoun rates as FCR increases. A possible Usage-based explanation for this is that although these forms share a high FCR, and therefore undergo the same qualitative influence from Switch Reference and Preceding Pronoun, they still differ in their *overall frequency* in the favoring context. For this reason, those with bigger exemplar clouds of tokens occurring in favoring contexts (i.e. those with greater absolute frequency of occurrence in favoring contexts) should have higher pronoun use than those with smaller exemplar clouds. The forms in the high-FCR group have very low overall frequencies ($N_{tokens} = 1,342, M_{raw\ frequency} = 6.5$), therefore it is possible that they do not have large enough exemplar clouds for sensitivity to FCR to manifest in a way that would impact their pronoun rates. The lack of positive correlation between the log frequencies and pronoun rates of high-FCR forms strongly indicates that overall frequency is still an important component in the realization of contextual frequency effects.

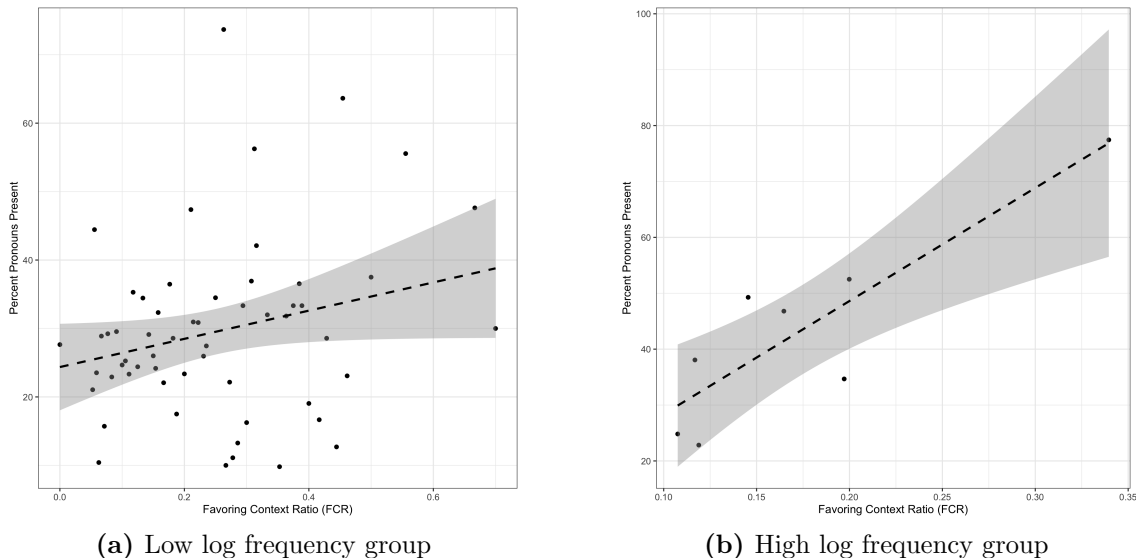


Figure 5.5: Pronoun rate by FCR for log-controlled groups

When we look at the low log frequency and high log frequency groups, we see that the predicted FCR effect – as FCR increases, pronoun rate increases – is borne

out in the high log frequency group. The correlation statistic shows a strong positive correlation between FCR and pronoun rate ($r(13, 157) = 0.88, p < 0.001$). In contrast, the correlation between FCR and pronoun rate for the low log frequency group is lower ($r(7, 030) = 0.04, p = 0.001$). Although still a significant effect, the FCR effect is much weaker for the infrequent forms. These results show that overall frequency impacts the effects of FCR, and when we control for frequency (by binning forms into groups based on log frequency), the effects of FCR are more in line with what we would expect. Figure 5-5 visualizes this: pronoun rate is trending in a positive direction as FCR increases from left to right along the x-axis for the high-log-frequency and low-log-frequency groups.

VERB	FCR	% PRONOUNS	<i>N</i> VERBS
<i>tengo</i> ‘I have’	0.11	24.8	2,111
<i>tenía</i> ‘I/he/she/you had’	0.12	38.1	1,387
<i>estoy</i> ‘I am’	0.12	22.8	1,489
<i>es</i> ‘he/she/you are’	0.15	49.3	1,100
<i>estaba</i> ‘I/he/she/you was/were’	0.16	46.8	1,421
<i>digo</i> ‘I say’	0.2	34.7	1,471
<i>creo</i> ‘I believe’	0.2	52.5	2,952
<i>sabes</i> ‘you know’	0.34	77.4	1,228

Table 5.5: Descriptive information for forms in the high log frequency group organized by FCR

More detailed information about the high-log-frequency forms that are presented in Figure 5-5b is described in Table 5.5. The table is organized from lowest FCR to highest FCR, and it shows that as FCR increases, the pronoun rates for the forms in the high log frequency group generally increase. Of these forms, *tengo* ‘I have’ has the lowest FCR and the second-lowest pronoun rate. The form *sabes* ‘you know’ has

the highest FCR at 0.34, and it has the highest pronoun rate (77.4%).

We now turn to mixed effects model results to investigate the extent to which FCR interacts with Log Frequency. As with interactions explored in the previous chapter, two mixed effects logistic regression models were run. The first model included FCR and Log Frequency as fixed effects and Verb as a random effect. The second model included FCR and Log Frequency as fixed effects and as an interaction. The interaction model also included Verb as a random effect. Results from the interaction model are presented in Table 5.6. The marginal R^2 for this model is 0.007, and the conditional R^2 is 0.200.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.477	0.037	-39.943	< 0.001***
FCR	0.980	0.093	10.571	< 0.001***
Log Frequency	0.111	0.021	5.384	< 0.001***
FCR*Log Frequency	-0.285	0.096	-2.973	0.003

Table 5.6: Interaction between FCR and Log Frequency; Model configuration: `glmer(Pronoun ~ FCR*Log Frequency + (1 | Verb))`; Reference level: 0.00 FCR, 0.00 Log Frequency, 0.00 FCR*0.00 Log Frequency

An ANOVA was run to determine the extent to which the regular model and the interaction model are significantly different. ANOVA results reveal that these models are significantly different ($p = 0.003$). Further, ANOVA results indicate that the interaction model is the more explanatory model with a lower AIC than the regular model (102,283 and 102,290, respectively). These results confirm what Figure 5-5 revealed: there is an overall effect of frequency that impacts the effect of FCR. The effect is not very strong, but it nevertheless supports the UBG notion that overall frequency of use impacts the robustness of mental representations of forms. The FCR

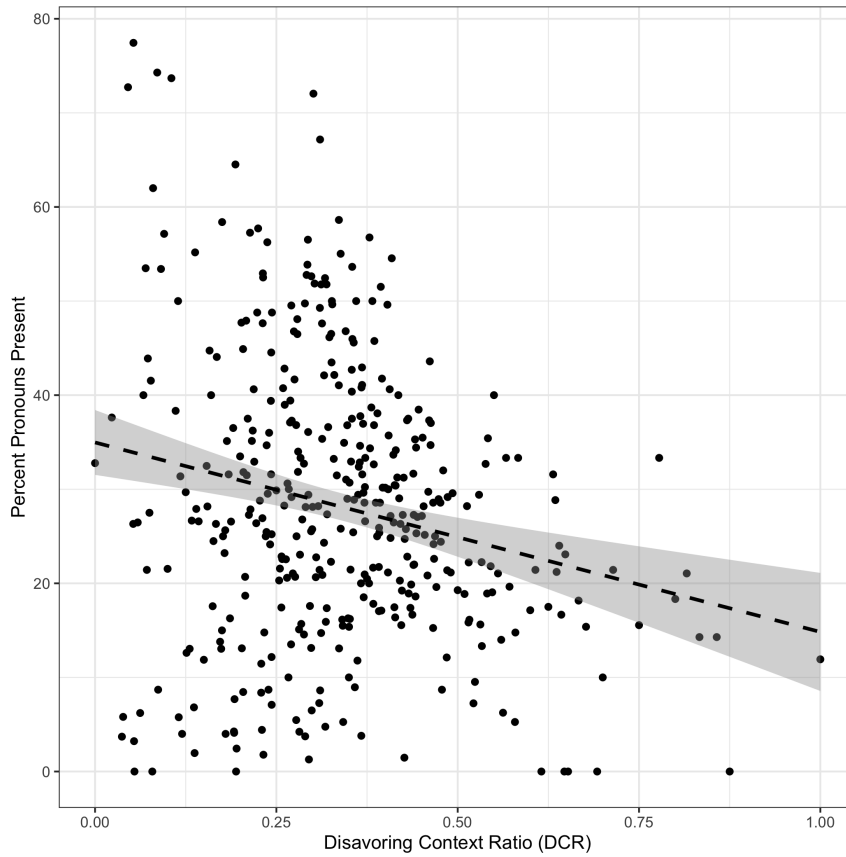


Figure 5-6: Disfavoring Context Ratio and percent pronouns present

of a given form only impacts pronoun use when the form reaches a certain overall frequency.

5.6 Disfavoring Context Ratio

For the final contextual frequency metric in the present study, we turn to DCR. The DCR is the number of times a verb appears in the pronoun disfavoring context (with the same referent and the preceding pronoun absent) relative to the overall frequency of the verb. We would predict that as DCR increases, pronoun rates will also decrease. This is exactly what we see. Figure 5-6 plots the pronoun rates for each DCR, and it shows a clear decrease in overt pronoun rates as DCR increases.

A Pearson's correlation statistic was run to statistically investigate the relationship presented in Figure 5.6. The correlation included pronoun rates and DCR. Results suggest a significant negative correlation between DCR and pronoun rates ($r(395) = -0.211, p < 0.001$). However, given the findings in the previous section on FCR, the same question arises: Does the DCR, which is behaving as expected, also depend on overall frequency?

As was done with the FCR, the dataset was divided into four groups to investigate a potential impact of overall frequency on the DCR. Two groups are based on DCR (one low DCR group between 0.1 and 0.2 and one high DCR between 0.4 and 0.5) and two are based on log frequency (one low log frequency and one high log frequency). Correlation tests that considered pronoun rate and log frequency of each finite form were run for the low- and high-DCR groups. I also ran a correlation statistic to test the relationship between DCR and pronoun rate when holding log frequency constant. As a reminder, holding DCR constant allows for a direct investigation of log frequency, and binning forms into low and high frequency permits a frequency-controlled test of the DCR. Correlation results are shown in Table 5.7.

GROUPED BY	GROUP	CORRELATION
Log Frequency	High Log Freq. (between 7.0 and 8.0)	$-0.78, p < 0.001$
	Low Log Freq. (between 2.0 and 3.0)	$-0.20, p < 0.001$
DCR	High DCR (between 0.4 and 0.5)	$0.15, p < 0.001$
	Low DCR (between 0.1 and 0.2)	$-0.038, p = 0.01$

Table 5.7: Summary of correlations between Log frequency and pronoun rate for low DCR and high DCR groups and correlations between DCR and pronoun rate for low log frequency and high log frequency groups.

The correlation results for the low-DCR group show a significant negative correlation between log frequency and pronoun rate ($r(4,527) = -0.038, p = 0.01$). In

other words, as log frequency increases, pronoun rate decreases. This is a frequency effect that we would not expect, given that this data show a direct effect of frequency on pronoun rate in the opposite direction. When we look at the correlation between log frequency and pronoun rate for the high-DCR group, we see quite different results. Correlation results indicate a positive correlation between log frequency and pronoun rates for high-DCR forms ($r(16, 209) = 0.15, p < 0.001$). Although this result is trending in the expected direction given an overall effect of frequency, it is actually against the expected effect of DCR (lower pronoun rates as DCR increases). As with the FCR results, these findings suggest that, although these forms are equally impacted by the DCR, they differ in their overall frequencies in the disfavoring context. This means that those with greater overall frequency of occurrence (which corresponds to larger exemplar clouds containing increased tokens of occurring in disfavoring contexts) should correspond to lower pronoun use than those with lower overall frequency. The forms in the low-DCR group have very low frequencies ($N_{tokens} = 4, 527, M_{raw\ frequency} = 45.2$), therefore they do not have large enough exemplar clouds for sensitivity to DCR to manifest in a way that would impact their pronoun rates. The lack of negative correlation between the log frequencies and pronoun rates of high-DCR forms reaffirms that overall frequency is playing a role in the effects of DCR.

Now let's turn to the high and low log frequency groups. When we look at these groups, we see the expected DCR effect – that higher DCR corresponds to lower pronoun rates – is borne out in both groups, but to different degrees. Results from the correlation statistic indicate a strong negative correlation between DCR and pronoun rate for the high log frequency group ($r(14, 367) = -0.78, p < 0.001$). In addition, the correlation between DCR and pronoun rate for the low log frequency group is significant ($r(6, 903) = -0.20, p < 0.001$). The DCR effect is much weaker for the low

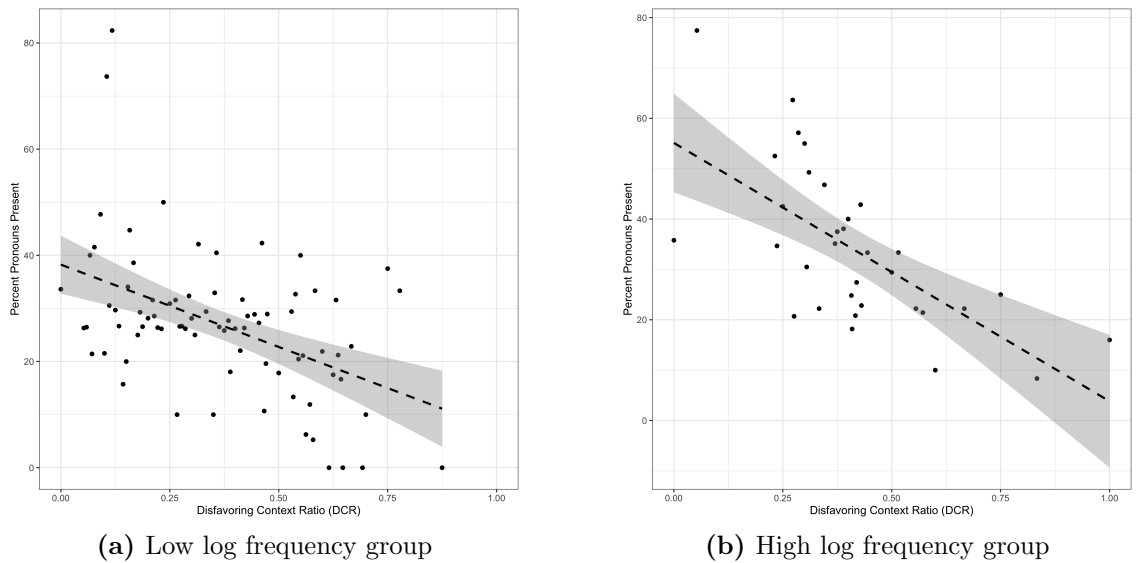


Figure 5.7: Pronoun rate by DCR for log-controlled groups

log frequency group than for the high log frequency group. Similar to the findings from FCR, these results indicate that when we control for overall frequency, DCR impacts pronoun production as predicted. A visual comparison of these findings is presented in Figure 5.7. As the figure shows, pronoun rate decreases as DCR increases for both groups, but the strength of the effect of DCR is lower for the low log frequency group (5.7a), since the slope of the smoothing line is less steep. This figure supports what was shown in the previous section on FCR: contextual frequency effects are themselves frequency-dependent.

Two mixed effects logistic regression models were run to explore the potential for an interaction between DCR and Log Frequency. The first model includes pronoun presence/absence as the response variable, DCR and Log Frequency as main effects and Verb as a random effect. The second model has the same configuration, except that it also includes DCR and Log Frequency as an interaction. Model results for the interaction model are presented in Table 5.8.

Interaction model results show that there is no interaction between DCR and

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.873	0.038	-22.856	< 0.001***
DCR	-1.472	0.090	-16.322	< 0.001***
Log Frequency	0.049	0.025	1.946	0.052
DCR*Log Frequency	0.086	0.070	1.235	0.217

Table 5.8: Interaction between DCR and Log Frequency; Model configuration: `glmer(Pronoun ~ DCR*Log Frequency + (1 | Verb))`; Reference level: 0.00 DCR, 0.00 Log Frequency, 0.00 DCR*0.00 Log Frequency

Log Frequency. The marginal R^2 for the interaction model (0.019) is lower than the marginal R^2 for the regular model (0.021). Further, ANOVA results indicate that the two models are not significantly different ($p = 0.22$). These results are interesting given that the correlation statistics in Table 5.7 indicate significant correlations between binned groups. These results suggest that DCR does not depend on overall frequency as FCR does.

Interestingly, the results presented here differ from those of FCR: (1) there is a smaller gap between the correlations for the low log frequency group and the high log frequency group for DCR, and (2) the DCR interaction model shows no significant interaction between log frequency and DCR. This difference is quite curious if we predict contextual ratios to have equal impact on the pronominal tendencies of verbs. However, as was uncovered in Section 5.1, the factor values for Switch Reference and Preceding Pronoun – which are the values that combine to create these contextual variables – do not have equal frequencies of occurrence or equal impact on pronoun use. Preceding Pronoun seems to be the stronger of the two variables, but a pronoun-favoring context with the preceding pronoun present can only impact pronoun use so much in a speech community with an overall pronoun rate of 31.6%. Not only does this

low overall pronoun rate speak to the tendency for Spanish speakers to favor non-use, but it also speaks to the low rate at which Preceding Pronoun is pronoun favoring. As discussed in Section 5.1, in a corpus where 31.6% of all possible sites of pronominal variation include a pronoun, only 31.6% of potential sites of variation are primed with a pronoun. In contrast, 68.4% of all possible sites of pronominal variation are primed with no pronoun. Therefore, pronoun-disfavoring contexts occur much more frequently in discourse in Spanish. Spanish speakers' proclivity to not use pronouns emerges from and is reinforced by these pronoun-disfavoring contexts.

Above all, these results show that frequency effects cannot be fully understood without disentangling overall frequency from contextual frequency. Contextual frequency measures cannot simply be substituted for lexical frequency measures to provide increased clarity as was predicted based on previous research. Instead, holding one frequency constant can allow for the exploration of potential non-linear frequency effects. With this approach, we are able to uncover frequency effects for frequent forms that are in-line with what is expected: increased frequency in (dis)favoring contexts impacts pronoun use.

Chapter 6

Discussion & Conclusions

The goal of this dissertation was to improve our understanding of the impact of frequency on morphosyntactic variation. The specific research questions that were asked in this thesis are: (1) Does replication with a larger, different data set produce results in line with those of Erker and Guy (2012)?; (2) Does incorporating contextual frequency metrics provide clearer results?; (3) What are the theoretical implications of incorporating insights from Usage-Based Grammar into Variationist Sociolinguistic research on Spanish subject pronoun production?; (4) Can the notion of a “prefab” be integrated into Variationist research?.

The following sections consider these questions further. Section 6.1 summarizes the replication results which correspond to the first research question. Section 6.2 reflects on the contextual frequency results. Section 6.3 investigates the theoretical implications of incorporating UBG insights into sociolinguistic research. Section 6.4 wrangles with the UBG notion of a “prefab” and investigates its role in sociolinguistic research. Then, sections 6.5 identify and discuss some additional considerations for expanding our investigation of contextual frequency in future research. This chapter concludes with a summary of the major takeaways from this study.

6.1 The Nature of the Replication Results

Overall, the results presented in Chapter 4 demonstrate that a larger, different data set *does* produce results in line with those of Erker and Guy (2012). The replication

portion of this thesis reveals not only an amplification effect of frequency for several linguistic conditioning factors (consistent with Erker and Guy (2012)), but also casts doubt on any main effect of overall frequency on pronoun presence/absence (though correlation statistics are significant). Not only do the figures reveal clear amplification effects of frequency for the majority of the linguistic constraints investigated (Morphological Regularity, Person/Number, Semantic Content, Switch Reference), but the significant interactions between these and discrete frequency provide statistical support for this modulating effect. The replication also recapitulates main effects for a series of linguistic predictors that have been shown to impact pronoun production: Person/Number, TMA, Semantic Category, and Switch Reference.

A more in-depth analysis of the pronominal tendencies of the top 32 most-frequent forms and inconsistent model results supports what is plotted in the raw frequency and log frequency plots: there are clear differences in verb rates for these forms, and frequency is shown to have no stand-alone effect in multiple mixed effects models. These findings provide the strongest support of the amplification effects of frequency on Spanish SPP variation since Erker and Guy (2012). The lack of a clear direct effect of frequency, although consistent with Erker and Guy (2012) and other replications, (Linford and Shin, 2013; Posio, 2015; Shin, 2016), nevertheless contradicts the findings of other previous replication attempts (Bayley et al., 2013; Posio, 2013; Rivas et al., 2018; Lease et al., 2022). Moreover, these findings motivate the second portion of the study, which expands upon Erker and Guy (2012) with the addition of contextual frequencies.

One question that emerges when considering the investigation of lexical frequency is the extent to which the increased variability in pronoun rates at higher frequencies is an artefact of individual verb forms. To further explore the relationship between overall frequency and pronoun use, Figure 6-1 plots the pronoun rates for each log

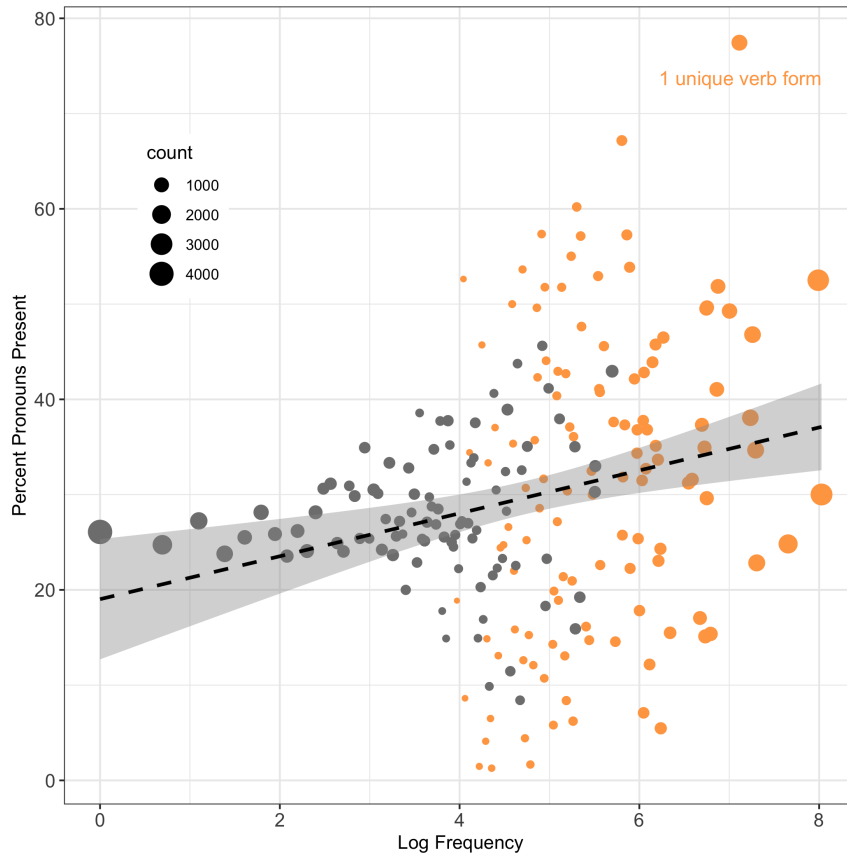


Figure 6·1: Percent pronouns present within each log frequency (accounting for token and type frequency)

frequency, with two additional parameters: (1) points colored orange consist of a single verb form and points colored grey consist of multiple verb forms; and (2) the size of each point corresponds to the number of observations at a given log frequency value, i.e. larger points correspond to more observations.

Figure 6·1 shows that the increase in variability begins to emerge prior to the shift to unique forms. As discussed in Chapter 4, once forms reach a log frequency of 4, the consistent positive correlation between log frequency and pronoun rate seemingly disappears and there is an immense increase in variability of pronoun rates. This figure shows that grey points (which consist of multiple verb forms) show increased variability between a log frequency of 4 and 6. By accounting for the points that

correspond to unique verb forms, the figure shows that this increased variability is not wholly an artefact of the behavior of singular verb forms. Instead, the figure shows that the positive relationship between pronoun and log frequency seems to arise in clusters of forms with identical log frequencies until a certain frequency is reached.

Figure 6·1 reinforces the findings from the replication analysis in Chapter 4, which clearly indicate that overall frequency is an inadequate predictor of pronoun use. How then can we account for the significant positive correlation between overall frequency and pronoun rates? One possible answer for this is that this positive correlation is merely an artifact of the overall low input probability of pronoun use in Spanish and the increased diversification of rates. As previously mentioned, the true effect of increased frequency is a great variability in pronoun rates. More frequent forms are more pronominally differentiated. When increased variation in rates comes into contact with a relatively low input probability (31.6% in the current study), a positive correlation such as the one discussed in Chapter 4 is perhaps guaranteed. This is due to the fact that there are more degrees of freedom (i.e. more available values) above the average rate than below it, given the possible rate of values for pronoun production ranges from 0 to 100. In other words, there is more space above the regression line than below within which diversification in rates may manifest. The very fact that there is increased dispersion in rates above the regression line in Figure 6·1 drives its slope upwards. This hypothesis allows for us to make predictions about the relationship between overall frequency and pronoun rates for other languages. For languages with overall low rates of pronoun use like Spanish, we would expect similar correlations between lexical frequency and pronoun rate. However, for languages with high overall pronoun rates (e.g. Brazilian Portuguese or Swabian German that show pronoun rates around 70-80%), we would predict inverse correlations between lexical

frequency and pronoun use such that increased frequency corresponds to decreased pronoun use. Future research should investigate the nature of this hypothesis cross-linguistically.

We now turn to a discussion of contextual frequency metrics.

6.2 The Incorporation of Contextual Frequency Metrics

The contextual frequency metrics do provide increased clarity on the nature of frequency effects in morphosyntactic variation to an extent. First, results show that the predicted pronoun-favoring and pronoun-disfavoring contexts, are, in fact, pronoun favoring and pronoun disfavoring. The initial general exploration of the four Switch Reference/Preceding Pronoun contexts also reveals that these contexts occur at quite different frequencies. The pronoun-disfavoring context (same referent/preceding pronoun absent), for example, occurs nearly twice as often as the pronoun-favoring context (different referent/preceding pronoun present). Further, the log contextual frequencies are highly correlated to log frequency, indicating that they are not, in and of themselves, sufficiently different measures for investigating contextual frequency effects.

Though the ratio-based metrics were better suited to investigate contextual frequency effects on Spanish variable pronoun use, they nevertheless provided some challenges. The linear fit of the initial FCR figure (Figure 5-3) suggested that FCR impacted pronoun rates in the opposite direction than was expected, and a Pearson's correlation statistic returned a negative, albeit insignificant, correlation between FCR and pronoun rate. A non-linear investigation of FCR provided much needed clarity and indicated a strong positive correlation between FCR and pronoun rate for high frequency forms. Model results also revealed an interaction between FCR and overall frequency. Similar methods were employed for DCR, and revealed significant cor-

relations between DCR and pronoun rate in the expected direction. The effects of FCR and DCR show that, although contextual frequency is important, overall frequency also plays a role in the emergence of sensitivity to relevant contexts. That is to say that for a (dis)favoring context to impact overall pronominal tendencies of a verb form, that verb form needs to be sufficiently frequent (i.e. with a large enough exemplar cloud).

This analysis highlights the difficulty in teasing apart overall frequency effects from contextual frequency effects, and underscores the importance of ratio-based contextual frequency metrics. Regardless of these challenges, the results provide strong evidence of the emergent nature of language knowledge. Contexts are important and are stored as details for each experienced finite form; but overall frequency is also important, since the size of an exemplar cloud (as suggested by verb frequency) modulates the effect of contextual frequency.

6.3 The Theoretical Implications of Incorporating UBG Insights into Variationist Research

In regards to the theoretical implications of incorporating UBG insights into Variationist research, there are multiple points to consider. First, it is clear that the traditional, binary “social”/“linguistic” divide for constraints is not ideal when considering the cognitively-motivated contextual frequency metrics put forth in this dissertation. One potential path forward would be to readjust the conceptualization of the current constraints as a three-way division that includes social, linguistic, and cognitive constraint categories. As discussed earlier in the dissertation, this notion is not entirely new¹. Tamminga et al. (2016) pushes for a similar adjustment to how variationists

¹Labov (2011) consists of an entire volume on cognitive and cultural factors. However, Labov’s definition of *Cognitive* is different from what is being proposed here. He defines cognitive factors in terms of language users’ perception of culturally-specific forms.

divide constraints, arguing for a new set of constraints grouped under what they call “p-conditioning”. Whether referred to as “p-conditioning” or “cognitive” constraints, the inclusion of a third category can ease tensions between traditional linguistic constraints, which typically only impact the sociolinguistic variable directly, and the cognitive constraints such as lexical and contextual frequency, which seem to impact all other linguistic constraints in addition to directly affecting the response variable.

In addition to the challenge of integrating contextual frequencies into the existing organization of binary constraints, it is also unclear how social aspects of linguistic experiences as outlined by UBG relate to social constraints as they are defined in variationist research. Exemplar theories have made clear that some social information is stored as part of the rich memory of each linguistic experience, but the specific nature of that social information is yet to be seen. Some studies have shown that perception and production of certain forms are facilitated when interlocutors are members of a social group with particular ties to that form (e.g. women more likely to use the word *children* than other genders) (Kim, 2011; Walker and Hay, 2011; Hay et al., 2019). This effect potentially signals some of the social information that is coded as part of each experienced exemplar. Usage-based grammarians have also suggested that the social information that is retained varies from language user to language user, since individual linguistic experiences and salience of social factors differ greatly across individuals. This is potentially at odds with the nature of social constraints in Variationist research, which present orderly heterogeneity, or variation that is exquisitely structured, at the community level (though of course exceptions exist within every community). As part of their investigation, Erker and Guy (2012) test the uniformity of the amplification effect of frequency across two dialects of Spanish: Dominican and Mexican. They find no difference in the effect of frequency for each group: discrete frequency potentiated linguistic constraints regardless of

dialect. The lack of any dialect effect suggests that, at the very least, region of origin is not a social factor that impacts frequency effects on variable pronoun use.

6.4 Testing for Pronominal Prefabrication

In addition to questions of contextual frequency, the current study addresses multiple open questions regarding UBG and how its insights may fit within Variationist research. One instance of this is the exemplar-theoretic notion of a ‘prefab’. Bybee (2010, p. 35) defines a prefab as “any conventionalized multi-word expression”, giving examples such as *X drives me crazy/mad* in American and British English, respectively. Bybee (2010) explains that expressions with extremely high frequencies, such as *X drives me crazy/mad*, can develop such robust mental representations that they develop separate exemplar clouds altogether – further facilitating their use. This description of a prefab suggests that a specific high-frequency variant of a form, such as *X drives me crazy/mad*, could actually be considered different or separate from other, less-frequent constructions of *X drives me Y*.

At the surface, it is unclear how this definition fits with the Labovian conception of a sociolinguistic variable, which as a reminder is defined as two or more ways of “saying the same thing” (Labov, 1972b, p. 188). In variationist studies of pronoun use, alternating between using a pronoun and not using a pronoun is considered an example of a sociolinguistic variable, since there are identical truth conditions with and without an overt pronoun. Based on the definition of a prefab, a Bybeeian explanation for highly frequent pro-drop forms might be that these verbs are “prefabs”, or separate forms, and therefore not variation at all. Let’s consider *tú sabes* and *sabes* ‘you know’ as an example. *Tú sabes* with a pronoun is generally much higher in frequency than its pronoun-free counterpart, appearing in the present corpus 951 times. Its high frequency might suggest that it is conventionalized, which, from a UBG per-

spective, could mean it is an entirely separate form from *sabes*, which appears without a pronoun only 277 times out of the 1,228 total number of occurrences. If these two forms are indeed represented as separate constructions and not variations of a single form, then these forms would not fall in line with the definition of a sociolinguistic variable, and should be excluded from variationist research. However, it is unclear how we determine the threshold at which we decide that these are truly separate prefabs that cannot be compared. This is especially challenging since *tú sabes* and *sabes* (and other tokens of Spanish pronoun variation) consist of two verb forms that differ only in the presence or absence of a pronoun, overlap significantly in their distribution, and have essentially equal truth conditions. Similarly, in English, “bathroom” and “restroom” are clearly independent lexical items that have related if not identical meanings and near identical distributions, but they are still in competition with each other.

The UBG theory of a prefab presents clear problems of integration for Variationist Sociolinguistic research. On purely theoretical grounds, there is no way to reconcile prefabs and sociolinguistic variables. As described above, they simply amount to different (and incompatible) ideas. However, since the previous chapter demonstrated that the current dataset is sensitive to contextual frequency in addition to other linguistic constraints, the present dissertation is now in a position to examine potential prefabs through quantitative exploration. An analysis like the one presented here can shed some light on the issue. To do this, I will consider three high-frequency forms from the current thesis that differ in their pronoun rates: (1) *sabes* ‘you know’, which has the highest pronoun rate of all high-frequency forms (77.4%); (2) *creo* ‘I believe’, which has a mid-range, but still relatively high pronoun rate (52.5%); and, (3) *son* ‘they are’, which has the lowest pronoun rate of all high-frequency forms (15.4%).

If these three highly frequent forms are in fact becoming conventionalized, they

should show reduced sensitivity to other linguistic constraints. To test this, I ran three mixed effects logistic regression models: one for each of the three verbs. These models included pronoun presence/absence as the response variable, switch reference and preceding pronoun as main effects², and speaker as a random effect³. To compare, I created three randomly sampled data sets from the whole corpus (excluding tokens of the possible prefabs) and ran mixed effects models with the same configurations on those subsets. Results are summarized in Table 6.1.

Results show a reduction of sensitivity to linguistic constraints that is predicted for highly frequent forms that are possibly becoming conventionalized. Table 6.1 presents the R^2 results for the three prefab models and three models with a random sampling of verbs. For all three verbs, we see that the marginal R^2 is much higher for the mixed effects regression models that considered a random sample of the same size as each verb. For instance, the *sabes* model is a logistic mixed effects model that considers switch reference and Preceding Pronoun as main effects and speaker as a random effect for all 1,228 instances of *sabes*. As shown in the table, the marginal R^2 for the random sample model is much higher than the R^2 for the *sabes* model ($R_m^2 = 0.051$ and $R_m^2 = 0.019$, respectively). The difference in marginal R^2 shows that *sabes* has less sensitivity to the linguistic constraints switch reference and Preceding Pronoun when compared to the random sample of other verbs. The difference in conditional R^2 shows increased variation is accounted for when we include speaker as a random effect for *sabes* – which suggests that some of this reduced sensitivity is speaker-dependant. These results suggest that instead of an online decision to use a

²Since these are all individual forms, they do not vary in person/number, TMA, morphological regularity, or semantic category of the verb. This leaves only Switch Reference and Preceding Pronoun as potential linguistic constraints on the pronoun production for these verbs.

³Verb was not included as a random effect for any of the models presented in Table 6.1, since the three prefab models only contain one verb. Instead, speaker was included as a random effect to rule out any type one errors and to account for the fact that *tú sabes* ‘you know’ as a prefab has been connected to specific regions and therefore could vary speaker to speaker.

	VERB ONLY	RANDOM SAMPLE
Sabes		
Count	1,228	1,228
Pronoun Rate	77.4%	29.7%
Marginal R^2	0.019	0.051
Conditional R^2	0.679	0.129
Creo		
Count	2,952	2,952
Pronoun Rate	52.5%	30.8%
Marginal R^2	0.008	0.056
Conditional R^2	0.430	0.144
Son		
Count	891	891
Pronoun Rate	15.4%	31.3%
Marginal R^2	0.048	0.087
Conditional R^2	0.171	0.127

Table 6.1: Summary of high frequency forms and their randomly sampled equivalents; Model configurations: `glmer(Pronoun ~ Switch Reference + Preceding Pronoun + (1 | Speaker))`

pronoun or not, based on the probability shaping power of conditioning factors, *sabes* has a pronominal inertia or bias that in and of itself contributes to the likelihood of pronoun presence/absence. In other words, *sabes* is a verb that favors pronoun use, independent of the online contextual constraints it has in a given utterance. This trend holds across all three verb/random sample pairs: the *creo* model shows lower marginal R^2 and higher conditional R^2 than its random sample counterpart; and the *son* model also has a lower marginal R^2 and higher conditional R^2 than its randomly sampled equivalent.

These findings indicate that when speakers use one of these three highly frequent forms, they are considering the fact that the verb is *sabes*, *creo*, or *son* and not another less-frequent verb. The extremely high frequency of use conventionalizes the pronominal tendencies of the verb such that the verb matters more than whether it is occurring in a switch in referent/preceding pronoun absent context, for instance. However, it is unclear how these forms developed their individual pronominal biases to begin with. This becomes especially challenging when we consider that these high frequency forms, which show reduced online sensitivity to Switch Reference and Preceding Pronoun, are also simultaneously sensitive to the FCR and DCR (per the results in Chapter 5), which are defined based on Switch Reference and Preceding Pronoun. Lease et al. (2022) offer a potential explanation for this asymmetry: frequent occurrence in specific constraining contexts reinforce the connection between a verb form and a pronoun (or lack thereof). In their example (*yo creo*), they point to the fact that it is first person singular, which, as previously stated, is known to favor pronoun use. Given this accumulation of experiences and the evidence of a contextual frequency threshold demonstrated in Chapter 5, it is possible that at a certain threshold, highly frequent forms have accumulated sufficient contextual information that the online context is no longer as important. This idea is not entirely new. Travis and Torres Cacoullos (2021) refer this phenomenon as “lexical effects”, or the cumulative effects of a language user’s experience with a form. Other studies have demonstrated that lexical effects can often hold more weight than online effects (Poplack, 1992; Cacoullos and Walker, 2009).

Overall, these findings suggest that prefabricated utterances do show decreased sensitivity to linguistic (and likely social) constraints. However, it is important to note that variation still exists for each of the three frequent forms investigated here. As discussed above, *sabes* occurs 227 times without a pronoun. *Creo* and *son* also ap-

pear in their less-frequent forms (*creo* and *ellos/ellas son*), showing variation in their production as well. How might a usage-based approach account for this variation? Are less-frequent variants of *tú sabes* (namely, *sabes*) stored in a separate exemplar cloud? In addition to evidence of variation, which poses a problem for prefabs for Usage-Based theory, *sabes* still presented some sensitivity to linguistic constraints in the direction that is expected. Although much less explanatory than the random sample model ($ps < 0.001$), the *sabes* model showed a significant decrease in pronoun likelihood for tokens with reference continuity ($\beta = -0.92, p = 0.009^{**}$) and a significant increase in pronoun likelihood for tokens with preceding pronoun present ($\beta = 0.74, p = 0.003^{**}$). *Creo* and *son* did not show significant main effects for both constraints, and instead each had a significant main effect for only one constraint: Preceding Pronoun was significant for *creo* ($\beta = 0.47, p < 0.001^{***}$), and Switch Reference was significant for *son* ($\beta = -0.83, p < 0.001^{***}$). These findings are curious, and cannot be attributed to differences in overall frequency, since *creo* is more frequent than *sabes*, while *son* is less frequent than *sabes*.

Although not directly related to the question of prefabrication, the difference in Conditional R^2 across individual verbs, which suggests differences in by-speaker variability, is particularly interesting. *Sabes* has the highest conditional R^2 ($R_c^2 = 0.679$) compared to *creo* ($R_c^2 = 0.430$) and *son* ($R_c^2 = 0.171$), suggesting that there is more by-speaker variability in pronoun use for *sabes*. This is particularly interesting since previous research that has investigated these highly-frequent mental activity verbs has discussed regional preferences for phrases such as *tú sabes*. Travis and Torres Cacoulios (2021) explain that the phrase *tú sabes* is much less frequent in their study of Spanish in Cali, Colombia, while studies of other Spanish varieties, such as Mexican/Mexican-American (Bayley et al., 2013) and Puerto Rican (Claes, 2011), indicate that *tú sabes* is the conventionalized form. It would be interesting

to explore questions of region of origin alongside questions of prefabrication like this one, to see if incorporating regional information increases clarity.

6.5 Further Considerations

There are multiple additional factors that could have been investigated in the present study. These include different methods for investigating frequency, the addition of social factors into the investigation, and others. Discussion on all of these are detailed in this section.

6.5.1 Different methodologies for frequency

One overarching concern that should be thoughtfully addressed in future work is the operationalization of frequency. In the studies cited in the dissertation alone, there are numerous different methods for measuring frequency. Some of these include categorical frequency at the 1% threshold, rank frequency, absolute frequency within the corpus, absolute frequency in an external corpus, log frequency, and others. The methodological challenges of defining frequency are extensive, since at minimum researchers must ensure that what is measured is an accurate reflection of speakers' frequency of use, determine where the distinction lies between frequent and infrequent within such a measurement, and account for inevitable differences in frequency due to the Zipfian distribution of words. These are all aspects that were considered in the present dissertation, but ultimately, decisions surrounding overall frequency were made based on the methodologies of Erker and Guy (2012) in order to replicate as closely as possible.

Nevertheless, there are some factors pertaining to frequency that could have been investigated and should be considered more closely in future research. One such frequency metric that could have been considered in the current thesis is rank frequency. Instead of measuring frequency as a count, which is the approach utilized

in the current dissertation, measuring frequency based on a verb form's rank in the corpus could eliminate some of the lexical frequency issues found in the present study. Erker and Guy (2012) did find comparable results between correlation statistics that included rank frequency as it correlates to pronoun use and raw frequency as it correlates to pronoun use. However, investigating rank frequency in a corpus of this magnitude could provide more clear insights. Additionally, rank frequency could be used as the threshold for determining discrete frequency labels, i.e. labelling the 25 most frequent words as "frequent" instead of using a percentage. This method has been illuminating in previous work: Erker (2011) found increased /s/-lenition in the 250 most frequent words in his corpus of Dominican Spanish. Further, since previous studies have defined categorically "frequent" forms according to vastly different counts, a rank frequency could remove some of these between-study differences and streamline our understanding of frequency in variation.

Another aspect of this study that could have been done differently is the method for choosing the threshold for discrete frequency: the current study simply retained the 1% cutoff from Erker and Guy (2012). Due to the sheer magnitude of the OZC-BSC Corpus, the 1% threshold does produce qualitatively different results than those of Erker and Guy (2012). The mean raw frequencies for forms labelled as "frequent" and "infrequent" in the current study are 1,903.4 and 231.6, while the mean raw frequencies for forms labelled as "frequent" and "infrequent" in Erker and Guy (2012) are 108.1 and 8.2. Overall, these share a similarly large disparity between mean raw frequency for each pair. However, the mean raw frequency for frequent forms in Erker and Guy's data is thirteen times greater than the mean raw frequency for infrequent forms. In contrast, the mean raw frequency for infrequent forms in the current study is eight times larger than the mean raw frequency for infrequent forms. This would be one reason to consider moving the threshold for being labelled "frequent" to below 1%

for the OZC-BSC data. This would yield a larger number of “frequent” forms, while subsequently driving the mean raw frequency for frequent and infrequent forms down. It is possible that a different threshold could better or more consistently illuminate lexical frequency effects in future research.

Additionally, the current study did not investigate the extent to which word length or phonetic length of the verb impacted pronoun production. It is possible that longer forms disfavor an overt pronoun in an effort to be more efficient or produce an utterance more quickly. This was not accounted for in the current study, nor was it investigated after data collection. It is also possible that contexts that consider other linguistic predictors (such as morphological regularity or TMA) provide stronger frequency effects. For this reason, it would be fruitful for future research to investigate different combinations of linguistic constraints when investigating how context-based frequency affects variable pronoun use.

6.5.2 Investigating social factors

The dissertation discusses at length the reasons for not including social factors into the current study. However, with a variationist investigation, one does question the extent to which social factors could impact the results presented here. This is especially the case with a corpus of such magnitude. It is unclear whether sex, age, or class would impact pronoun rates directly in this corpus, or indirectly through frequency or another metric, or not at all. It is possible that social differences manifest only when they are accounted for with differences in usage. In other words, it is possible that speakers that differ along social categories have different frequencies of use, and investigating social factors without considering contextual frequency conceals social effects.

It would be particularly interesting to investigate whether the difference between pronoun rates for pronoun-favoring and -disfavoring contexts holds across different

regions of origins. There is consistent research that shows that the level of sensitivity to certain predictors differs based on the geographic location or regional origin of speakers. In other words, effect sizes of well-known linguistic and social constraints on variation have been shown to differ across different speech communities. Investigating whether this holds for the pronoun-favoring different referent/preceding pronoun present context could provide insight into whether this is true for cognitive constraints as well.

Another issue that is investigated in variationist research that was not addressed in the present study is the issue of language contact. That is to say that it is not clear how questions related to language contact would be best integrated into exemplar-theoretic models or into variationist models that incorporate Frequency and other exemplar-theoretic variables. Some previous research has shown that Spanish-English Bilinguals produce higher overt pronoun rates in Spanish than Spanish monolinguals matched for region of origin (Otheguy and Zentella, 2012). How might this be addressed in a Bybee model for mental language representation? If contact with English is influencing Spanish subject pronoun production, this might suggest that exemplars from both languages and their relevant contextual information are stored in a single, united exemplar cloud such that SPPs experienced in English contribute to a generally high pronoun rate across languages. This is an empirical question that has serious implications for studies of contextual frequency.

6.6 Other Avenues for Future Analyses

In addition to the considerations for frequency and the analysis of social factors described above, there are also other avenues to consider for further analyses. One such analysis would be a deeper exploration into binning the OZC-SBC Corpus. For example, it would be very interesting to see the results of an analysis of just plural

tokens. Specifically, conducting a comparison of frequent vs. infrequent forms in a subset of the data that includes only plural tokens may yield illuminating results. It is possible that plurals, which are overall less frequent at least in the context of a sociolinguistic interview, will present different patterns related to overall frequency and contextual frequency. This is not something that was explored in the current study, but it is something that may be worth considering in future analyses.

6.7 Conclusions

In all, this dissertation disentangles the complex relationship that exists between lexical frequency, contextual frequency, and Spanish subject pronoun variation. From this investigation comes two overarching conclusions: First, speakers are keeping track of the contextual properties of verbs; Second, lexical frequency and contextual frequency must be considered in tandem. Lexical frequency clearly interacts with other linguistic constraints (albeit inconsistently), and it amplifies the effects of those linguistic constraints almost unanimously. Further, contextual frequency provides increased clarity into the relationship between frequency and pronoun variation. However, understanding the effects of contextual frequency cannot be done without considering lexical frequency, too. Contextual frequency effects are contingent on lexical frequency. Mental representations of verb forms of lower frequency, then, do not contain robust enough information to present sensitivity to the contextual properties of those forms. In contrast, mental representations of extremely frequent forms contain robust enough contextual information that they develop their own pronominal inertia, which outweigh online constraints.

Overall, results indicate that linguistic knowledge (e.g. the sensitivity to linguistic constraints) is emergent from usage, since speakers are sensitive to contextual *and* lexical frequency in subject pronoun variation. Ultimately, this dissertation provides

strong insight into the role of frequency effects in morphosyntactic variation and the impact of frequency on mental grammar. This dissertation highlights a need for increased awareness of frequency (contextual and lexical) in future investigations of morphosyntactic variation.

Appendix A

Appendix

A.1 Speaker Demographics

The demographic information of the speakers in the OZC-BSC Corpus is provided in this section of the Appendix. The demographic information described here includes factors such as age, sex, region of origin, country of origin, percent of life in the U.S. (PLUS), and age of arrival. The definitions and numeric breakdowns of each of these social variables are presented below.

Table A.1 presents further breakdown of speakers' Age, and Sex in each corpus and overall in the joint corpus. The average age for speakers overall is 34.4 years old, while the average age for Boston (33.2) is slightly younger than that for NYC (34.8). There are more female speakers than male speakers in the corpus ($N_{female} = 117$; $N_{male} = 104$). The Boston corpus has 8 more female speakers ($N_{female} = 44$) than male speakers ($N_{male} = 36$), and the NYC corpus has 5 more female speakers ($N_{female} = 73$) than male speakers ($N_{male} = 68$).

A.1.1 Regional Origin

Table 3.1 presents the speakers' Regional Origin and Country of Origin by Sex for each corpus and overall. Regional Origin has five categories: Andean, Caribbean, Central, European, and Mixed. Speakers are coded as "Andean" if they are from Colombia, Ecuador, Paraguay, Peru, and parts of Venezuela. Speakers are from the "Caribbean" if they are from Cuba, Dominican Republic, Puerto Rico, and parts of Venezuela.

	Boston	NYC	OZC-BSC
	(N = 80)	(N = 141)	(N = 221)
Age			
Mean (SD)	33.2 (13.0)	34.8 (13.3)	34.4 (13.2)
Median [Min, Max]	31.0 [18.0, 73.0]	31.0 [12.0, 80.0]	31.0 [12.0, 80.0]
Sex			
Female	44 (55%)	73 (51.8%)	117 (52.9%)
Male	36 (45%)	68 (48.2%)	104 (47.1%)

Table A.1: Age and Sex for Speakers in the Boston Corpus, NYC Corpus, and OZC-BSC joint corpus.

Speakers are coded as “Central” if they are from El Salvador, Guatemala, Honduras, Mexico, and Nicaragua. Speakers are coded as European if they are from Spain. Finally, speakers are coded as “Mixed” if their parents are from countries in different regions (i.e. El Salvador and Dominican Republic). The largest group of speakers overall are from the Caribbean ($N_{Caribbean} = 97$), followed by Andean ($N_{Andean} = 65$), Central ($N_{Central} = 56$), European ($N_{European} = 2$), and Mixed ($N_{Mixed} = 1$). As for the breakdown of regional origin by corpus, the Central region is the most common region for speakers in the Boston corpus ($N_{Central} = 33$), followed by the Caribbean ($N_{Caribbean} = 25$), then Andean ($N_{Andean} = 19$), European ($N_{European} = 2$), and Mixed ($N_{Mixed} = 1$). In the NYC corpus, the largest group of speakers originates in the Caribbean ($N_{Caribbean} = 72$), then the Andean region ($N_{Andean} = 46$), followed by Central ($N_{Central} = 23$). There are no European or Mixed speakers in the NYC corpus.

Participants’ Country of Origin was coded based on speakers’ self-reported re-

	Boston		NYC		OZC-BSC	
	Female	Male	Female	Male	Female	Male
	(N = 44)	(N = 36)	(N = 73)	(N = 68)	(N = 117)	(N = 104)
Regional Origin						
Andean	10 (22.7%)	9 (25.0%)	27 (37.0%)	19 (27.9%)	37 (31.6%)	28 (26.9%)
Caribbean	14 (31.8%)	11 (30.6%)	35 (47.9%)	37 (54.4%)	49 (41.9%)	48 (46.2%)
Central	20 (45.5%)	13 (36.1%)	11 (15.1%)	12 (17.6%)	31 (26.5%)	25 (24.0%)
European	0	2 (5.6%)	0	0	0	2 (1.9%)
Mixed	0	1 (2.8%)	0	0	0	1 (1.0%)
Country of Origin						
Colombia	5 (11.4%)	5 (13.9%)	11 (15.1%)	11 (16.2%)	16 (13.7%)	16 (15.4%)
Colombia/Peru	1 (2.3%)	0	0	0	1 (0.9%)	0
Cuba	0	0	14 (19.2%)	10 (14.7%)	14 (12.0%)	10 (9.6%)
Dominican Republic	5 (11.4%)	5 (13.9%)	11 (15.1%)	14 (20.6%)	16 (13.7%)	19 (18.3%)
Ecuador	1 (2.3%)	0	16 (21.9%)	8 (11.8%)	17 (14.5%)	8 (7.7%)
El Salvador	11 (25.0%)	7 (19.4%)	0	0	11 (9.4%)	7 (6.7%)
El Salvador/D.R.	0	1 (2.8%)	0	0	0	1 (1.0%)
Spain	0	2 (5.6%)	0	0	0	2 (1.9%)
Guatemala	1 (2.3%)	2 (5.6%)	0	0	1 (0.9%)	2 (1.9%)
Honduras	1 (2.3%)	0	0	0	1 (0.9%)	0
Mexico	6 (13.6%)	4 (11.1%)	11 (15.1%)	12 (17.6%)	17 (14.5%)	16 (15.4%)
Nicaragua	1 (2.3%)	0	0	0	1 (0.9%)	0
Paraguay	1 (2.3%)	0	0	0	1 (0.9%)	0
Puerto Rico	6 (13.6%)	5 (13.9%)	10 (13.7%)	13 (19.1%)	16 (13.7%)	18 (17.3%)
Peru	2 (4.5%)	3 (8.3%)	0	0	2 (1.7%)	3 (2.9%)
Venezuela	3 (6.8%)	2 (5.6%)	0	0	3 (2.6%)	2 (1.9%)

Table A.2: Regional origin and Country of origin for Speakers in the Boston Corpus, NYC Corpus, and Overall.

sponses. Participants in the present study come from fourteen different countries, and four participants are from two countries of origin (Colombia & Peru or El Salvador & Dominican Republic). The seven most common countries of origin for the speakers in the current study are the Dominican Republic ($N_{DR} = 35$), Puerto Rico ($N_{PR} = 34$), Mexico ($N_{ME} = 33$), Colombia ($N_{CO} = 32$), Ecuador ($N_{EC} = 25$), Cuba ($N_{CU} = 24$), and El Salvador ($N_{EL} = 18$). These seven countries represent 90.1% of all speakers in the data, and the seven remaining countries of origin represent the remaining 20 speakers. More specific break downs by sex for each corpus are presented in Table 3.1.

A.1.2 Arrival to the U.S.

Participants were recruited in either New York City, NY or Boston, MA. However, when each speaker arrived in these northeastern cities varied. Age of Arrival (AOA) is a quasi-continuous variable that was coded based on speakers' self-reported age of arrival to NYC or Boston (See Table A.3). The average AOA for the Boston corpus is slightly younger than the NYC ($Mean_{AOA} = 16.3$ & 19.4 , respectively). The overall average AOA of the joint corpus is 18.3 years old. The oldest AOA in the Boston corpus is 52 years old, while the oldest AOA in NYC is 70 years old.

Each speaker's Percent of Life in the U.S. (PLUS) was calculated by subtracting the speaker's AOA from their age at the time of the interview and then dividing that number by their age and multiplying by 100. The result is a semi-continuous variable that can range from 0.00% to 100.0%. For example, if a 20-year-old speaker has lived in the U.S. for 5 years, their PLUS is 25% ($(5/20) * 100$). Boston speakers have higher average PLUS than NYC speakers (53.4% and 45.8%, respectively). This is unsurprising given that Boston speakers have a younger average AOA than NYC speakers. Overall, the average PLUS in the joint corpus is 48.6%. The median, minimum and maximum PLUS for each corpus are presented in Table A.3.

	Boston	NYC	OZC-BSC
	(N=80)	(N=141)	(N=221)
AOA			
Mean (SD)	16.3 (13.4)	19.4 (14.2)	18.3 (14.0)
Median [Min, Max]	17.0 [0, 52.0]	20.0 [0, 70.0]	19.0 [0, 70.0]
PLUS			
Mean (SD)	53.4 (34.5)	45.8 (32.4)	48.6 (33.3)
Median [Min, Max]	50.0 [0, 100]	37.9 [0, 100]	44.2 [0, 100]

Table A.3: Age of arrival (age) and PLUS for Speakers in the Boston Corpus, NYC Corpus, and OZC-BSC Corpus.

References

- Abramowicz, L. (2007). Sociolinguistics Meets Exemplar Theory: Frequency and Recency Effects in (ing). *University of Pennsylvania Working Papers in Linguistics*, 13(2):27–37.
- Abreu, L. (2009). *Spanish subject personal pronoun use by monolinguals, bilinguals, and second language learners*. PhD thesis, University of Florida.
- Abreu, L. (2012). Subject pronoun expression and priming effects among bilingual speakers of Puerto Rican Spanish. In Geeslin, K. and Díaz-Campos, M., editors, *Selected Proceedings of the 14th Hispanic Linguistics Symposium*, pages 1–8, Somerville. Cascadilla.
- Albright, A., Andrade, A., and Hayes, B. (2000). Segmental Environments of Spanish Diphthongization. *UCLA Working Papers in Linguistics (Papers in Phonology 5)*, pages 117–151.
- Albright, A. and Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.
- Alfaraz, G. (2015). Variation of overt and null subject pronouns in the Spanish of Santo Domingo. In Carvalho, A., Orozco, R., and Lapidus Shin, N., editors, *Subject pronoun expression in Spanish: A cross-dialectal perspective*, pages 3–16. Georgetown University Press, Washington, D.C.
- Anderson, H. (2013). La influencia de la persona gramatical sobre la expresión del pronombre sujeto en el español del sur de Arizona [The influence of grammatical person on subject pronoun expression in Southern Arizona Spanish]. *Divergencias: Revista de estudios lingüísticos y literarios*, 11(1):35–47.
- Ávila-Jiménez, B. (1995). A sociolinguistic analysis of a change in progress: Pronominal overtiness in Puerto Rican Spanish. *Cornell Working Papers in Linguistics*, (13):25–47.
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.
- Bayley, R., Cardenas, N. L., Treviño Schouten, B., and Velez Salas, C. M. (2012). Spanish dialect contact in San Antonio, Texas: An exploratory study. In Geeslin,

- K. and Holland, C., editors, *Selected proceedings of the 14th Hispanic Linguistics Symposium*, pages 48–60, Sommerville. Cascadilla Proceedings.
- Bayley, R., Greer, K., and Holland, C. (2013). Lexical Frequency and Syntactic Variation: A Test of a Linguistic Hypothesis. *U. Penn Working Papers in Linguistics*, 19:21–30.
- Bayley, R., Greer, K. A., and Holland, C. L. (2017). Lexical frequency and morphosyntactic variation: Evidence from U.S. Spanish. *Spanish in Context*, 14(3):413–439.
- Bayley, R. and Pease-Alvarez, L. (1997). Null pronoun variation in Mexican-descent children’s narrative discourse. *Language variation and change*, 9(3):349–371.
- Bentivoglio, P. (1980). *Why Canto and Not yo Canto? The Problem of First-Person Subject Pronoun in Spoken Venezuelan Spanish*. PhD thesis, University of California, Los Angeles, CA.
- Bentivoglio, P. (1987). Los sujetos pronominales de primera persona en el habla de Caracas. Technical report, Universidad Central de Venezuela, Caracas.
- Berkenfield, C. (2001). The role of frequency in the realization of English that*. In Bybee, J. and Hopper, P. J., editors, *Frequency and the Emergence of Linguistic Structure*, pages 281–308. John Benjamins Publishing Company.
- Bessett, R. M. (2018). Testing English influence on first person singular “yo” subject pronoun expression in Sonoran Spanish. In MacDonald, J. E., editor, *Contemporary trends in Hispanic and Lusophone linguistics*, pages 355–372. John Benjamins Publishing Company.
- Bongiovanni, S. (2014). “¿Tomas [pepsi], [peksi] or [pesi]?”: A variationist sociolinguistic analysis of Spanish syllable coda stops. *IULC Working Papers in Linguistics*, pages 43–61.
- Bouchard, M. E. (2018). Subject Pronoun Expression in Santomean Portuguese. *Journal of Portuguese Linguistics*, 17:1–29.
- Brown, E. L. (2015). The role of discourse context frequency in phonological variation: A usage-based approach to bilingual speech production. *International Journal of Bilingualism*, 19(4):387–406.
- Brown, E. L. and Raymond, W. D. (2012). How discourse context shapes the lexicon: Explaining the distribution of Spanish f-/ h-words. *Spanish and Portuguese Faculty Contributions*, 3.

- Brown, E. L., Raymond, W. D., Brown, E. K., and File-Muriel, R. J. (2021). Lexically specific accumulation in memory of word and segment speech rates. *Corpus Linguistics and Linguistic Theory*, 17(3):625–651.
- Brown, E. L. and Shin, N. (2022). Acquisition of cumulative conditioning effects on words: Spanish-speaking children’s [subject pronoun + verb] usage. *First Language*, 42(3):361–382.
- Brown, E. L. and Torres Cacoulios, R. (2003). Spanish /s/ reduction: A different story from beginning (initial) to end (final). In Nunez-Cedeno, R., Lopez, L., and Cameron, R., editors, *A Romance perspective on language knowledge and use*, pages 21–38. John Benjamins, Amsterdam/Philadelphia.
- Bush, N. (2001). Frequency effects and word-boundary palatalization in English. Amsterdam: John Benjamins. In Bybee, J. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pages 255–280. John Benjamins, Amsterdam.
- Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins, Amsterdam and Philadelphia.
- Bybee, J. (1998). A functionalist approach to grammar. *Evolution of Communication*, 2(2):249–278.
- Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge.
- Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation & Change*, 14:261–290.
- Bybee, J. (2006). From usage to grammar: the mind’s response to repetition. *Language*, 82:711–733.
- Bybee, J. (2010). *Language, usage, and cognition*. Cambridge University Press.
- Bybee, J. and Eddington, D. (2006). A usage-based approach to Spanish verbs of ‘becoming’. *Language*, 82:323–355.
- Cacoulios, R. and Walker, J. (2009). The present of the English future: Grammatical variation and collocations in discourse. *Language*, pages 321–354.
- Cacoulios, T. and Travis, C. E. (2019). Variationist typology: shared probabilistic constraints across (non-)null subject languages*. Technical report.
- Callen, M. C. and Miller, K. (2022). Linguistic Variation in the Acquisition of Morphosyntax: Variable Object Marking in the Speech of Mexican Children and Their Caregivers. *Language Learning and Development*, 18(3):310–323.

- Cameron, R. (1992). *Pronominal and null subject variation in Spanish: Constraints, dialects, and functional compensation*. University of Pennsylvania, Philadelphia, PA.
- Cameron, R. and Flores-Ferrán, N. (2004). Perseveration of subject expression across regional dialects of Spanish. *Spanish in Context*, 1(1):41–65.
- Carvalho, A. and Bessett, R. (2015). Subject pronoun expression in Spanish in contact with Portuguese. In Carvalho, A., Orozco, R., and Lapidus Shin, N., editors, *Subject pronoun expression in Spanish: A cross-dialectal perspective*, pages 143–168. Georgetown University Press, Washington, D.C.
- Carvalho, A., Orozco, R., and Shin, N. (2015). *Subject pronoun expression in Spanish: A cross-dialectal perspective*. Georgetown University Press.
- Carvalho, A. M. and Child, M. (2011). Subject pronoun expression in a variety of Spanish in contact with Portuguese. In Michnowicz, J. and Dodsworth, R., editors, *Selected Proceedings of the 5th Workshop on Spanish Sociolinguistics*, pages 14–25, Somerville. Cascadilla Proceedings Project.
- Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.
- Chomsky, N. (1965). *Aspects of a Theory of Syntax*. The MIT Press, Cambridge.
- Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York.
- Claes, J. (2011). ¿ Constituyen las Antillas y el Caribe continental una sola zona dialectal?: Datos de la variable expresión del sujeto pronominal en San Juan de Puerto Rico y. *Spanish in context*, 8(2):191–212.
- Díaz-Campos, M. (2005). The role of frequency in the study of phonological variation: A usage-based analysis of /r/ deletion in Venezuelan Spanish. *Paper presented at the Hispanic Linguistics Symposium & the Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*.
- Díaz-Campos, M. and Gradoville, M. (2011). An Analysis of Frequency as a Factor Contributing to the Diffusion of Variable Phenomena: Evidence from Spanish Data. In Ortiz-López, L. A., editor, *Selected proceedings of the 13th Hispanic Linguistics Symposium*, pages 224–238, Somerville. Cascadilla Proceedings Project.
- Dionne, D. and Coppock, E. (2022). Complexity vs. salience of alternatives in implicature: A cross-linguistic investigation. *Glossa Psycholinguistics*, 1(1).
- Enriquez, E. V. (1984). *El pronombre personal sujeto en la lengua española hablada en Madrid*. Consejo Superior de Investigaciones Científicas, Madrid.

- Erker, D. (2011). *An acoustic sociolinguistic analysis of variable coda /s/ production in the Spanish of New York City*. PhD thesis, New York University, New York.
- Erker, D. (2022). How social salience can illuminate the outcomes of linguistic contact: Data from Spanish in Boston. In Guy, G. R. and Beaman, K., editors, *The coherence of linguistic communities: Orderly Heterogeneity and Social Meaning*, pages 145–162. Routledge Studies in Sociolinguistics.
- Erker, D. and Guy, G. (2012). The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language*, 88(3):526–557.
- Escalante, C. (2016). Tracking a shifting target: A longitudinal exploration of s-weakening among L2 speakers of coastal Ecuadorian Spanish. In *Eighth International Workshop of Spanish Sociolinguistics*, San Juan, Puerto Rico.
- Flores-Ferrán, N. (2002). *Subject personal pronouns in Spanish narratives in Puerto Ricans in New York City: A sociolinguistic perspective*, volume 2. Lincom Europa.
- Flores-Ferrán, N. (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.
- Givón, T. (1983). Topic continuity in discourse: The functional domain of switch reference. *Switch reference and universal grammar*, 51(82).
- Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago University Press, Chicago.
- Goldberg, A. (2006). *Constructions at Work*. Oxford University Press.
- Gudmestad, A., House, L., and Geeslin, K. L. (2013). What a Bayesian Analysis Can Do for SLA: New Tools for the Sociolinguistic Study of Subject Expression in L2 Spanish. *Language Learning*, 63(3):371–399.
- Guy, G. R. (2014). Linking usage and grammar: Generative phonology, exemplar theory, and variable rules. *Lingua*, 142:57–65.
- Hay, J., Walker, A., Sanchez, K., and Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLoS ONE*, 14(2).
- Hay, J. B., Pierrehumbert, J. B., Walker, A. J., and LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, 139:83–91.
- Hochberg, J. (1986). Functional compensation for /s/ deletion in Puerto Rican Spanish. *Language*, 62(3):609–621.

- Holmquist, J. (2012). Frequency rates and constraints on subject personal pronoun expression: Findings from the Puerto Rican Highlands. *Language Variation and Change*, (24):203–220.
- Hooper, J. B. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In Christie, W. M., editor, *Current progress in historical linguistics*, pages 96–105. North Holland, Amsterdam.
- Kanwit, M. and Geeslin, K. L. (2020). Sociolinguistic competence and interpreting variable structures in a second language. *Studies in Second Language Acquisition*, 42(4):775–799.
- Kim, J. (2011). Perceptual associations between word and speaker age. *Laboratory Phonology*, 7(1):1–22.
- Labov, W. (1963). The social motivation of a sound change. *Word*, 19(3):273–309.
- Labov, W. (1966). *The social stratification of English in New York City*. Cambridge University Press.
- Labov, W. (1972a). Sociolinguistic Patterns. *Conduct and Communication*.
- Labov, W. (1972b). Some Principles of Linguistic Methodology. *Language in Society*, 1(1):97–120.
- Labov, W. (1994). *Principles of Linguistic Change, Volume 1: Internal Factors*. Wiley-Blackwell, Malden, MA.
- Labov, W. (2001). *Principles of Linguistic Change, Volume 2: Social Factors*. Blackwell, Malden, MA.
- Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, 34(4):500–515.
- Labov, W. (2011). *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors*, volume 3. John Wiley & Sons.
- Lapidus, N. and Otheguy, R. (2005a). Contact induced change? Overt nonspecific *ellos* in Spanish in New York. In Sayahi, L. and Westmoreland, M., editors, *Selected proceedings of the Second Workshop on Selected proceedings of the Second Workshop on Spanish Sociolinguistics*, pages 67–75, Somerville, MA. Cascadilla Press.
- Lapidus, N. and Otheguy, R. (2005b). Overt nonspecific “*ellos*” in the Spanish of New York. *Spanish in Context*, 2:157–176.

- Lapidus Shin, N. and Erker, D. (2015). The emergence of structured variability in morphosyntax: Childhood acquisition of Spanish subject pronouns. In Carvalho, A. M., Orozco, R., and Shin, N. L., editors, *Subject pronoun expression in Spanish: A cross-dialectal perspective*. Georgetown University Press.
- Lastra, Y. and Butragueno, P. M. (2015). Subject pronoun expression in Oral Mexican Spanish. In Carvalho, A. M., Orozco, R., and Lapidus Shin, N., editors, *Subject pronoun expression in Spanish: A cross dialectal perspective*, chapter 3, pages 39–58. Georgetown University Press, Washington, D.C.
- Lease, S., Shin, N. L., and Bird-Brown, E. (2022). Community Norms and Lexical Frequency Shape U.S. Bilingual Children’s Subject Pronoun Expression. *Heritage Language Journal*, 19(1):1–29.
- Li, X. and Bayley, R. (2018). Lexical frequency and syntactic variation: Subject pronoun use in Mandarin Chinese. *Asia-Pacific Language Variation*, 4:135–160.
- Linford, B. and Shin, N. L. (2013). Lexical frequency effects on L2 Spanish subject pronoun expression. *Selected proceedings of the 16th Hispanic linguistics symposium*, pages 175–190.
- Michnowicz, J. (2015). Subject pronoun expression in contact with Maya in Yucatan Spanish. In Carvalho, A., Orozco, R., and Lapidus Shin, N., editors, *Subject pronoun expression in Spanish: A cross-dialectal perspective*, pages 101–119. Georgetown University Press, Washington, D.C.
- Miyajima, A. (2000). Spanish subject pronoun expression and verb semantics. *Sophia Linguistica*, 46/47:73–88.
- Myers, J. and Guy, G. R. (1997). Frequency effects in Variable Lexical Phonology. *University of Pennsylvania Working Papers in Linguistics*, 4(1):215–228.
- Myhill, J. (2005). Quantitative methods of discourse analysis. In Altmann, G. and Piotrowski, R. G., editors, *Quantitative linguistics: An international handbook*, pages 471–498. Mouton de Gruyter, Berlin.
- Orozco, R. (2015). Pronominal variation in Costeño Spanish. In Carvalho, A., Orozco, R., and Lapidus Shin, N., editors, *Subject Pronoun Expression in Spanish: A Cross-dialectal Perspective*, pages 17–37. Georgetown University Press, Washington, D.C.
- Orozco, R. (2018). *Spanish in Colombia and New York City: Language contact meets dialectal convergence*. IMPACT: Studies in Language and Society. John Benjamins Publishing Company, Amsterdam.

- Orozco, R. (2022). El efecto del verbo en la variación lingüística: Expresión de sujetos pronominales. In Ruiz Vásquez, N. F., editor, *Perspectivas actuales de la investigación en lingüística: entre tradición y modernidad*, pages 53–96. Instituto Caro y Cuervo, Bogotá, Colombia.
- Orozco, R. and Guy, G. (2008). El uso variable de los pronombres sujetos: ¿Qué pasa en la costa Caribe Colombiana? In Westmoreland, M. and Thomas, J. A., editors, *Selected Proceedings of the Fourth Workshop on Spanish Sociolinguistics*, pages 70–80, Somerville. Cascadilla Proceedings Project.
- Orozco, R. and Hurtado, L. M. (2021). A Variationist Study of Subject Pronoun Expression in Medellín, Colombia. *Languages*, 6(5):1–29.
- Otheguy, R. and Zentella, A. C. (2012). *Spanish in New York: Language contact, dialectal leveling, and structural continuity*. Oxford University Press.
- Otheguy, R., Zentella, A. C., and Livert, D. (2007). Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language*, pages 770–802.
- Padilla, L. (2021). First person singular subject pronoun expression in Equatoguinean Spanish. *Journal of Monolingual and Bilingual Speech*, 3(2):171–194.
- Padilla, L. V. K. (2020). *Subject Pronoun Expression in an L2-only Environment: The Case of Equatorial Guinea*. PhD thesis, Arizona State University.
- Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. and Hopper, P., editors, *Frequency effects and the emergence of linguistic structure*, pages 137–158. John Benjamins, Amsterdam.
- Pierrehumbert, J. B. (2002). Word-specific phonetics. In *Laboratory Phonology 7*, pages 101–140. De Gruyter Mouton.
- Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, 34(4):516–530.
- Poplack, S. (1992). The inherent variability of the French subjunctive. In Laeuffer, C. and Morgan, T., editors, *Theoretical Analyses in Romance Linguistics*, pages 235–263. Benjamins, Amsterdam.
- Poplack, S. (2001). Variability, frequency, and productivity in the irrealis domain of French. In Bybee, J. and Hopper, P., editors, *Frequency and the Emergence of Linguistic Structure*, page 405. John Benjamins, Amsterdam/Philadelphia.
- Posio, P. (2011). Spanish subject pronoun usage and verb semantics revisited: First and second person singular subject pronouns and focusing of attention in spoken Peninsular Spanish. *Journal of Pragmatics*, 43(3):777–798.

- Posio, P. (2013). The expression of first-person-singular subjects in spoken Peninsular Spanish and European Portuguese: Semantic roles and formulaic sequences. *Folia Linguistica*, 47(1):253–292.
- Posio, P. (2015). Subject pronoun usage in formulaic sequences. In *Subject pronoun expression in Spanish: A cross-dialectal perspective*, pages 59–78. Georgetown University Press.
- Prada Pérez, A. d. (2009). *Subject expression in Minorcan Spanish: Consequences of contact with Catalan*. PhD thesis, Penn State University.
- Prada Pérez, A. d. (2015). First person singular subject pronoun expression in Spanish in contact with Catalan. In Carvalho, A., Orozco, R., and Lapidus Shin, N., editors, *Subject pronoun expression in Spanish: A cross-dialectal perspective*, chapter 7, pages 121–142. Georgetown University Press, Washington, D.C.
- Purse, R., Fruehwald, J., and Tamminga, M. (2022). Frequency and morphological complexity in variation. *Glossa: a journal of general linguistics*, 7(1).
- Raymond, W. D. and Brown, E. L. (2012). Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In Gries, S. T. and Divjak, D., editors, *Frequency Effects in Language Learning and Processing*, volume 1, pages 35–52. De Gruyter Mouton, Berlin/Boston.
- Raymond, W. D., Brown, E. L., and Healy, A. F. (2016). Cumulative context effects and variant lexical representations: Word use and English final t/d deletion. *Language Variation and Change*, 28(2):175–202.
- Rivas, J., Brown, E. L., and Cortés-Torres, M. (2018). Variable subject pronominal expression in non-finite clauses: Implications for variant patterns and emergent contexts. *Lingua*, 215:27–39.
- Rodríguez-Ordóñez, I. (2022). The role of frequency in the acquisition of structured variation: The case of Basque ergativity. *International Journal of Bilingualism*, 26(5):656–672.
- Shin, N. L. (2013). Women as leaders of language change: A qualification from the bilingual perspective. In Carvalho, A. M. and Beaudrie, S., editors, *Selected Proceedings of the 6th Workshop on Spanish Sociolinguistics*, pages 135–147, Somerville. Cascadilla Proceedings Project.
- Shin, N. L. (2014). Grammatical complexification in Spanish in New York: 3sq pronoun expression and verbal ambiguity. *Language Variation and Change*, 26(3):303–330.

- Shin, N. L. (2016). Acquiring constraints on morphosyntactic variation: children's Spanish subject pronoun expression. *Journal of Child Language*, 43(4):914–947.
- Shin, N. L. and Otheguy, R. (2013). Social class and gender impacting change in bilingual settings: Spanish subject pronoun use in New York. *Language in Society*, 42(4):429–452.
- Silva-Corvalán, C. (1982). Subject Expression and Placement in Spoken Mexican-American Spanish. *Spanish in the United States: Sociolinguistic Aspects*, pages 93–120.
- Skinner, B. F. (1957). *Verbal Behavior*. Appleton-Century-Crofts, New York.
- Stanford, J. N. (2019). *New England English : Large-scale acoustic sociophonetics and dialectology*. Oxford University Press.
- Tamminga, M. (2014). Sound Change without Frequency Effects: Ramifications for Phonological Theory. In Santana-LaBarge, R. E., editor, *Proceedings of the 31st West Coast Conference on Formal Linguistics*, pages 457–465, Somerville. Cascadilla Proceedings Project.
- Tamminga, M., MacKenzie, L., and Embick, D. (2016). The dynamics of variation in individuals. *Linguistic variation*, 16(2):300–336.
- Todd, S., Pierrehumbert, J. B., and Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, 185:1–20.
- Torres Cacoullous, R. (2000). *Grammaticization, Synchronic Variation, and Language Contact: A study of Spanish progressive -ndo constructions*. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Torres Cacoullous, R. and Travis, C. E. (2011). Testing convergence via code-switching: priming and the structure of variable subject expression. *International Journal of Bilingualism*, 15(3):241–267.
- Torres Cacoullous, R. and Travis, C. E. (2015). Foundations for the Study of Subject Pronoun Expression in Spanish in Contact with English: Assessing Interlinguistic (Dis)similarity via Intralinguistic Variability.
- Torres Cacoullous, R. and Travis, C. E. (2018). *Bilingualism in the Community: Code-switching and grammars in contact*. Cambridge University Press.
- Travis, C. E. (2007). Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change*, 19(2):101–135.

- Travis, C. E. and Torres Cacoullos, R. (2012). What do subject pronouns do in discourse? Cognitive, mechanical, and constructional factors in variation. *Cognitive Linguistics*, 23(4):711–748.
- Travis, C. E. and Torres Cacoullos, R. (2021). Categories and frequency: Cognition verbs in Spanish subject expression. *Languages*, 6(3).
- Vidal Covas, L.-A. M. (2013). El uso variable de los pronombres sujetos en el castellano puertorriqueño hablado en Luisiana y Puerto Rico. *LSU's Masters Thesis*, 3876.
- Walker, A. and Hay, J. (2011). Congruence between ‘word age’ and ‘voice age’ facilitates lexical access. *Laboratory Phonology*, 1(2):219–237.
- Weinreich, U., Labov, W., and Herzog, M. I. (1968). Empirical Foundations for a Theory of Language Change. In Lehmann, W. P. and Malkiel, Y., editors, *Directions for Historical Linguistics*, pages 95–195. University of Texas Press, Austin, Texas.
- Wolfram, W. (1991). The Linguistic Variable: Fact and Fantasy. *American Speech*, 66(1):22–32.
- Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, 32:251–256.

CURRICULUM VITAE

