5-1-2012

SUBJECT EXPRESSION IN BRAZILIAN PORTUGUESE: CONSTRUCTION AND FREQUENCY EFFECTS

Agripino Silveira

Agripino de Souza Silveira Neto
Candidate

Linguistics
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Catherine E. Travis, Ph.D. , Chairperson

Joan Bybee, Ph.D.

Richard File-Muriel, Ph.D.

Rena Torres-Cacoullos, Ph.D.

Alexandra Aikenvald, Ph.D.

SUBJECT EXPRESSION IN BRAZILIAN PORTUGUESE:

CONSTRUCTION AND FREQUENCY EFFECTS

BY

AGRIPINO DE SOUZA SILVEIRA NETO

B.A. In English and Portuguese Languages and Literatures,
Federal University of Ceará – Brazil, 2000
M.A. in Portuguese, University of New Mexico, 2004

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Linguistics

The University of New Mexico
Albuquerque, New Mexico

May 2012

DEDICATION

To my parents, Lucilene and Eduardo

To my sister, Alexandrina

To my angel, Nick

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, mentor, and friend Catherine E. Travis for her encouragement, positivism, and most importantly, for her care in the several reviews this work went through, and for all her help throughout my career as a linguist at the University of New Mexico. Catherine has inspired me to become a researcher who strives to understand the subtleties of spoken language. I thank you. My gratitude is also extended to my dissertation committee, Joan Bybee, Richard File-Muriel, Rena Torres-Cacoullos and Alexandra Aikhenvald, whose mentoring have shaped my interpretation of language and linguistics.

To Frances Garcia who took me in and gave me the means to continue my work and persevere in my dreams. Thank you so much.

I am grateful to my parents, my sister, and Nick for their emotional support and for believing in me. I know it must have been hard for you all, and I cannot express how grateful I am.

I must also thank the Latin American and Iberian Institute for their financial support through the Title VI Doctoral Fellowship. I am also grateful to the Office of Graduate Studies for the two semester Dean's Fellowship. I also want to thank all the departments I worked for during all these years at the University of New Mexico, namely the Portuguese and Linguistics Departments and the Center for English Language and Culture (CELAC).

I am very grateful to Professor Lemos Monteiro who so kindly allowed me to digitize and use his corpus of spoken Portuguese (PORCUFORT), without which this work would have been completed.

Finally, I would like to thank all those that helped one way or another throughout my journey here. I know that a few words of thanks do not express my gratitude, but know that I will always be indebted to you. Thus, I would like to thank Manolisa Vasconcelos, Vládia Borges Cabral, Odirene Bezerra, Margo Milleret, Ricardo Paiva, Alexandro de Sousa, Valico Romualdo, Ana Christina Powell, Larry Smith and Paul Edmunds.

SUBJECT EXPRESSION IN BRAZILIAN PORTUGUESE:

CONSTRUCTION AND FREQUENCY EFFECTS

BY

AGRIPINO DE SOUZA SILVEIRA NETO

B.A. In English and Portuguese Languages and Literatures,
Federal University of Ceará – Brazil, 2000
M.A. in Portuguese, University of New Mexico, 2004
Ph.D., Linguistics, University of New Mexico, 2012

ABSTRACT

Brazilian Portuguese (henceforth BP) has for long been considered as a Null-subject language due to its variability in regards to subject expression (e.g. *Era bom porque eu diminuía de peso... era muito gordinha* 'That was good because then I could lose some weight… (I) was a bit chubby.' C33:179). Such variability has been attributed to the language's once rich inflectional system, and the reported increase in rate of subject expression has been seen as a result of changes to the system (Barbosa, Duarte, & Kato, 2005; Monteiro, 1994b; Negrão & Viotti, 2000). Moreover, there is agreement among several scholars that the variability can still be accounted for in terms of traditional factors such as emphasis, clarity, and ambiguity of the Tense, Aspect, and Mood (TAM) system. In this work, I demonstrate that, rather than an effect of such pragmatic factors as these, subject expression in BP is to a large degree an artifact of the frequency of use of certain constructions of different degrees of fixedness.

The analysis proposed here falls under the framework of usage-based linguistics in which grammar is believed to be shaped by discourse as speakers produce it online (Bybee, 2006). Thus, any linguistic pattern observed in speech is emergent and a result of repetition

(Bybee, 2006; Hopper, 1998). Therefore, it is believed that the patterns of subject expression found in the data are a result of the speaker's experiences with those patterns.

The data used for the study are drawn from the corpus of oral Portuguese as spoken by educated speakers from Fortaleza (PORCUFORT) (Monteiro, 1994a). The analysis is based on 8066 tokens of 1sg, 2sg, and animate 3sg subjects culled from three different registers (Conversations, Interviews, and Lectures) across three different age groups (22-35, 36-50, and over 51). These tokens are subjected to a number of multivariate analyses to identify the contexts that significantly contribute to the realization of pronominal subjects in these data.

The methodology employed in this study to analyze the data follows the tenets of the Comparative method in Variationist theory in that comparison across the different subjects allows us to identify the contexts that contribute to the overall pattern of pronominal subjects. Moreover, this analysis also takes into account the role of frequency and constructions in shaping the grammar of speakers.

These different analyses and approaches yield two major findings from this study, namely (1) that these three persons behave very differently in terms of their patterning with pronominal subjects, they show that there are different factor groups conditioning the realization of pronominal subjects and within these factor groups we see that the factors show different directions of effect depending on the person; (2) that high frequency verbs and constructions also behave differently in their distribution with pronominal subjects. In fact, their behavior is needs to be examined in isolation because some show regular patterning with pronominal subjects while others are realized without pronominal subjects.

TABLE OF CONTENTS

List of Tables

List of Figures

1 INTRODUCTION

In the context of functional linguistics, researchers are interested in analyzing language as it is produced by speakers for any purpose their linguistic production may serve. Within such a perspective linguists have moved from conceiving of grammar as an abstract arrangement of pre-determined rules, to a more concrete description of human processes that interact in the production, perception, and storage of language. Thus, Bybee contends that, in a theory based on language usage, the grammar has to be defined as "the cognitive organization of one's experience with language" (2006, p. 711). In this cognitive perspective, grammar is not seen as a static system, but rather as a structure that emerges from use (Hopper, 1998), especially as a result of communicative events that speakers perform on a daily basis. In short, the frequency with which words and structures occur together plays a role in shaping grammar.

Usage-based linguistics postulates that linguistic items and structures are gradient and highly affected by input – e.g. frequency among other factors (Bybee, 2001, p. 20). In this sense, frequency of input becomes rather important in establishing the relational connections within categories. As frequency of input increases, linguistic items become stronger and easier to access. Therefore, the storage of linguistic structures and lexical items is in part contingent on frequency effects. Storage is conceived not as a list of items but as a network of connections, which are strengthened between the items that share similar properties (Bybee, 1985).

When applied to syntax, usage-based linguistics started looking at constructions that have a tighter constituency, e.g. idioms (Kövecses & Szabo, 1996; Wray, 2000). These constructions are putatively accessed as single units; therefore, they are rendered non-compositional. It has been observed, however, that not only idioms can be interpreted in this

fashion, but any other kinds of fixed expressions that show a frequent rate of co-occurrence of their constituent parts. Bybee and Scheibman (1999), for example, show that the expression *I don't know* is accessed as a whole in certain environments, suggesting storage of the expression as a unit rather than it being derived by rule. This is evidenced in part by phonological reduction in the form, in particular the fact that the vowel [ə] in *don't* is reduced most when it occurs in the construction *I don't know*, as compared with other contexts.

It is under the umbrella of usage-based linguistics that this study intends to account for a much discussed issue in Brazilian Portuguese (henceforth BP), namely subject expression. The present study analyzes the patterns of pronominal expression for first person singular (1sg), second person singular (2sg), and third person singular (3sg) animate[1] subjects in declarative clauses based on a total of 8,066 tokens extracted from naturally occurring Brazilian Portuguese discourse. The working hypothesis that is showcased in this study lies on the premise that discourse is intrinsically connected to the grammar a speaker holds in their minds, that is, discourse not only shapes the grammar, but reinforces it as well. As stated earlier, frequency comes to play an important role in the way linguistic structures are stored, perceived, accessed, and produced by speakers. Thus, utterances are to some extent an artifact of the frequency in which they normally occur in discourse. With that in mind, it is proposed here that subject expression, as discourse in general, is also affected by frequency. It is hypothesized that certain forms and combinations of subjects and verbs tend to be more

[1] This study only takes into account animate 3sg referents because there are only two possible pronominal forms to refer to them, namely *ele* 'he' and *ela* 'she'. Inanimate referents can be expressed through several other forms, including the latter two. Moreover, excluding inanimate referents provides a more methodologically sound basis for comparison between the three persons.

frequent in discourse, and the use of expressed or unexpressed subjects is a product of the frequency of co-occurrence of these items. Throughout this study it will be demonstrated that frequency does play a role in the way structures emerge in discourse, consequently affecting the variability in subject expression found in BP.

1.1 Hypotheses

Brazilian Portuguese shows variable subject expression as (1) where the same speaker produces both expressed and unexpressed subjects. This phenomenon has been discussed extensively not only in the literature in Brazilian Portuguese (Barbosa, 1995; Barbosa, et al., 2005; Duarte, 1993, 2003; Mary Aizawa Kato, 1999, 2000; Lira, 1982; M. Modesto, 2000a; Paredes Silva, 2003; V. L. P. Silva, Santos, & Ribeiro, 2000; Silveira, 2008), but also in Spanish (Bentivoglio, 1987; Cameron, 1992, 1995; Cameron & Flores-Ferrán, 2003; Enríquez, 1984, 1986; Flores-Ferrán, 2002; Morales, 1997; Otheguy, Zentella, & Livert, 2007; Silva-Corvalán, 1982; Torres Cacoullos & Travis, 2010; Travis, 2005, 2007) and Italian (Rizzi, 1986). The research in subject expression has evoked both formal and functional explanations to try to understand the nature and conditioning of this variability, and although there is great diversity in the nature of the explanations, it is concurred among researchers in Brazilian Portuguese that this language is moving toward obligatory, non-variable subject expression.

(1) ***Ela*** *nasceu com um dedinho a mais do que...*
do lado da mão direita...
mas só aquele cotoquinho...
Ø fez cirurgia logo quando Ø era pequena mesmo.
'**She** was born with a bit of an extra finger…
on her right hand…
but just that tiny bit…
(she) had surgery when **(she)** was a child.'

(C34: 194-196)[2]

Taking into consideration the findings to date on the conditioning of the variability of pronominal and unexpressed subjects in BP, the main questions addressed in this dissertation are formally outlined as follows:

a. Based on the linguistic factors found in the literature to condition subject expression, which ones determine the realization of expressed or unexpressed subjects in these data in BP?
   i. It is expected that factors usually claimed in traditional analysis to have an effect on subject expression in BP, i.e. ambiguity of TAM, and change of referents no longer have an effect on subject expression in BP.
   ii. In terms of discourse organization and topic continuity, it is expected that more topically continuous referents will be realized as unexpressed subjects, whereas less continuous referents are anticipated to be realized pronominally based on research on subject expression in both Spanish and Portuguese (Ávila-Shah, 2000; Paredes Silva, 1993, 2003; V. L. P. d. Silva, 1996), as well as studies on information flow (cf. the papers in Givón, 1983c).
b. How are the three persons of speech, specifically 1sg, 2sg, and 3sg, different or similar to one another? Since the forms included in the analysis will not differ in terms of animacy.

[2] The letter refers to the register from which the example was extracted (C – Conversations; L – Lectures; I – Sociolinguistic interviews). The first number, to the left, refers to the transcript, and the second number refers to the lines in the transcript.

c. Several studies both in BP as well as in Spanish have shown that the person highly conditions the realization of pronominal subjects (Duarte, 1993, 2000, 2003; Otheguy, et al., 2007; Silva-Corvalán, 1982, 1994, 2001). Other studies have examined different persons of speech and showed different patterning for each of the persons, namely, for 1sg subjects Silveira (2008), Travis (2005, 2007), and Torres Cacoullos and Travis (2010), for 2sg subjects, Silva, Santos & Ribeiro (2000) and Faraco (1996) for 2sg subjects, and Silva (2003; 1996) for 3sg subjects. Thus, it is hypothesized here that each person will show distinct conditioning of the realization of pronominal subjects.

d. How do type and token frequency interact with subject expression?

   i. Both the frequency of the main verb and the frequency of co-occurrence of the main verb and a particular subject, i.e., bigram frequency (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009), are anticipated to have an effect in different ways:

      1. Frequently co-occurring collocations of subject and verb will be strongly resistant to variation, showing strong tendencies either with expressed or unexpressed subjects
      2. Verbs and subjects that do not co-occur as frequently, on the other hand, are assumed to be more prone to being realized with expressed subjects;

These hypotheses will guide us through the subsequent chapters in the analysis of subject expression in this data.

1.2 Outline of the Dissertation

The overall structure of this dissertation takes the form of 6 themed chapters, this introductory chapter and a conclusion. This introductory chapter lays out the hypotheses, objectives and the theoretical dimensions of this research, and it looks at the way this dissertation will be framed by usage-based linguistics analysis, supported by two of its components, namely the notions of frequency and of constructions.

Chapter two gives a review of the notion of subjects in Brazilian Portuguese and of the studies on subject expression in Brazilian Portuguese both through the formalist and the functionalist lenses. This chapter focuses on the notion of subjects in BP and how the variation between pronominal subjects and unexpressed subjects is obtained in the language.

The third chapter is concerned with the methodology that guides this dissertation. We examine the framework of Variationist Linguistics along with its core premise, the envelope of variation. Within this chapter, we also develop the hypotheses into operationalized factor groups that will later be subjected to statistical analyses to obtain the conditioning constraints that guide the realization of pronominal subjects in BP. In this chapter, we discuss the theoretical tenets of this theory, the contexts in which there is variation between pronominal and unexpressed subjects, the data used in this study, and the factor groups to be tested in the statistical analyses are also discussed.

Chapters four, five, six, and seven present the results of the statistical analyses conducted on the data as well as the discussion of the constructions that emerged from the data. The fourth chapter presents the results of an overall statistical analysis conducted on 8,066 tokens with all three persons combined. In this analysis seven of the eight factor groups included were selected as significant in the conditioning of pronominal expression

(**VERB CLASS**, **PERSON**, **TAM**, **MODAL**, **CLAUSE TYPE**, **DISCOURSE CONTINUITY**[3], and **POLARITY**). The discussion of these results suggests that the three persons are indeed behaving differently in their patterning of pronominal expression.

Chapter five analyzes the results of separate statistical analyses conducted on each person including the same set of factor groups, except for **PERSON**. From these analyses it is learned that each person is indeed conditioned differently by these factor groups in their realization with pronominal subjects. The differences are so stark that while one factor group may play a strong role in variant choice with one particular person, it may not be even selected as significant among the others, or more strikingly, it shows a different direction of effect. These finding reverberate in one of our hypotheses in that these persons should be analyzed separately.

Chapter five also demonstrates that within each analysis there is a strong lexical effect interacting with the different factor groups. These findings lead us to reconsider one of our hypotheses, namely that which predicted that the frequency of certain predicates would have an effect on the way pronominal subjects are realized. This hypothesis is further tested in chapter six where several statistical analyses are conducted with highly occurring verbs with 1sg, 2sg, and 3sg subjects excluded in each. What the results of these analyses demonstrate is that there is a pronounced difference in the way the data behaves when these highly frequent forms are removed from the statistical analysis.

The seventh and last results chapter tackles the lexical forms, or constructions, that were excluded from the analyses presented in chapter 6 to assess their effect on variant

[3] This refers to the discourse continuity of the subjects, namely the three persons examined in this work.

choice. In this chapter, I argue that pronominal expression is highly dependent on whether a form frequently occurs with it or not.

The conclusion draws upon the entire dissertation, tying up the various theoretical and empirical strands in order to understand more fully the conditioning of pronoun realization with these three persons of speech.

1.3 Usage-based Linguistics

In the context of functional linguistics, researchers are interested in analyzing language as it is produced by speakers for any purpose their linguistic production may serve. Within such a perspective, linguists have moved from conceiving of grammar as an abstract arrangement of pre-determined rules or ordered constraints, to a more concrete description of human processes that interact in the production, perception, and storage of language, and, crucially, this is not distinct from other cognitive properties. Thus, Bybee contends that, in a theory based on language usage, grammar has to be defined as "the cognitive organization of one's experience with language" (2006, p. 711). In this cognitive perspective, grammar is not seen as a static system, but rather as a structure that emerges from use (Hopper, 1998), especially as a result of communicative events that speakers perform on a daily basis.

Thus, a usage-based model assumes that a speaker's linguistic system is fundamentally grounded in 'usage events', i.e. "instances of the speaker's producing and understanding language" (Kemmer & Barlow, 2000, p. viii). These instances are the basis on which a speaker's linguistic system is formed, and they are essentially specific in nature[4].

[4] As will be discussed later in this section, a speaker's linguistic system consists of both the specific and the general. Specific instances of linguistic input and output are stored whole, while generalizations emerge from the similarities between several usage events.

Hence, the linguistic system is built up from such instances, only gradually abstracting more general representations. These representations can be of any level of linguistic analysis, e.g. phonemes, morphemes, constructions, etc. Such representations form what can be called the units of language, and these are not fixed but dynamic in nature; they are subject to reshaping due to use (Bybee, 2006; Kemmer & Barlow, 2000; Langacker, 2000).

In the usage-based model, linguistic units are seen as cognitive routines, i.e. recurrent patterns of mental, and ultimately neural, activation. Thus, a particular location in the brain is not postulated to store these units as is assumed in more traditional models (Jackendoff, 2004). This belief that units are not stored in one single location in the brain is in agreement with findings in psychology and the neurosciences regarding the lack of central processing units in the brain that directs mental operations (Kandel, Schwartz, & Jessell, 2000; MacWhinney, 2005; McClelland & Patterson, 2002). Instead, each neuron is its own processor and functions by activating or inhibiting links to other neurons. This is an important premise in connectionist models in that different candidates, i.e. different representations of the same linguistic unit, compete for activation and their output is the result of simultaneous constraint satisfaction rather than a rule-like process.

A usage-based model is dynamic in nature since linguistic structure is susceptible to usage effects. In this sense, then, linguistic structure is thought of as *emergent* (Hopper, 1998). That is, frequent usage events such as *não sei* 'I don't know' or *digamos* 'let's say' emerge and come to be stored as independent units in the lexicon/grammar. More abstract structures can emerge from commonalities across different usage events, too, such as [*vamos* V-INF] 'let's V-INF'. In this way, constructions or schemas come to be stored in the lexicon where constructions are "specific sequential units, often containing explicit morphological

material, which have at least one variable slot in which any member of a category may appear" (Bybee, 2002a, p. 6). Thus, the usage-based model is redundant (Kemmer & Barlow, 2000, p. ix), storing constructions alongside fully instantiated expressions which themselves have autonomous storage, i.e., separate lexical representation (Bybee, 1985), even though they could be arrived at by accessing the appropriate constructions.

As was previously mentioned, in the usage-based model, the speaker's linguistic system is comprised of both general and specific items. However, substantial importance is placed on the actual use of the linguistic system and the speaker's knowledge of this use, thus being considered a functionalist approach to linguistic analysis. Thus, it can be asserted that the usage-based model is a non-reductionist approach to linguistic analysis because it does not presuppose that linguistic forms are stored at different levels, rather it is assumed that linguistic units are stored whole in the speakers' minds.

The appropriateness of a non-reductionist approach is that both linguistic units as well as abstractions are stored in the speaker's mind, against Cartesian linguistics that implied the need for rules to generate the grammar of a language and any forms that deviated from these rules were stored in list form (Chomsky, 1965, 1991). Instead, what is stored in the speaker's mind is a series of psychological events. Over time, and through repetition, these events coalesce into routines that are easily accessible and reliably executed. Once structures achieve such an automated status in that they are manipulated as a pre-packaged assembly, they can be considered to form a linguistic unit (cf. Langacker, 2000, p. 3 inter alia).

The usage-based model also requires acknowledging the role of human cognition in the organization of grammar. Clearly, the emergence of constructions requires the human mind be able to categorize and generalize, or make abstractions. Additionally, the human

cognitive ability to track linguistic details must be incorporated into the model. With respect to frequency, lexical strength (Bybee, 1985) and resting activation rate (Jurafsky, 1996) are two similar theoretical constructs that have been invoked to capture the fact that the mental lexicon tracks how often a linguistic unit is used, i.e., its token frequency. Thus, a frequently used unit is weighted such that it is primed for future use. If the unit is highly frequent, it may even come to be entrenched (Travis & Silveira, 2009). That is, the unit undergoes chunking and automatization (Haiman, 1994) or forms a cognitive routine (Langacker, 2000). A result of entrenchment is that the form will often become fused to function as a single unit. Phonological fusion is one aspect of this phenomenon (Bybee, 2002c). Resistance to regularization and lexical split (Bybee, 1985) as well as greater lexical and grammatical idiomaticity (Erman & Warren, 2000) are other indicators of a unit's automated status. Such linguistic phenomena result because the routinized unit tends to lose its lexical connections to other similar forms and gains autonomous representation (Bybee, 1985). To illustrate this point, take the verb *saber* 'to know' which is most frequently realized in four constructions, i.e. *sei* '(I) know$_{\text{-PRESENT}}$', *não sei* '(I) don't know$_{\text{-PRESENT}}$' and *num sei* '(I) don't know$_{\text{-PRESENT}}$', *sabe* '(you) know$_{\text{-PRESENT}}$'. What can be seen here with these constructions is that they are independent from general syntactic patterns, namely general syntactic rules that govern overall realzation of subject expression. Instead, they have their own patterning of subject expression or lack thereof, and these appear to be representated individually in the speaker's minds.

Finally, once complex units are admitted into the usage-based lexicon, various processing mechanisms must also be included as part of the grammar because irrespective of how a unit came to be stored holistically, presumably it is accessed as a whole too.

Therefore, two types of processing have been proposed (Sinclair, 1991): the idiom principle and the open choice principle. The idiom-principle processing operates by accessing a store of "semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (Sinclair, 1991) while open choice processing represents "a more standard view of syntax in which syntactic composition makes reference only to syntactic categories, not lexical items" (Barlow & Kemmer, 2000). Although both processing strategies are recognized, the idiom principle is typically given preference in the usage-based model, not only for irregular and idiomatic expressions but for regular and non-idiomatic expression as well (Barlow & Kemmer, 2000; Bybee, 1995; Bybee & Cacoullos, 2009; Sinclair, 1991).

The distinction between the idiom and open choice principle can be seen in the way constructions will be analyzed in this study. As an example, let us consider the idiom TER + X + ANOS, 'be X years old', literally 'have X years', where X represents a number of years as illustrated in (2). While it might appear that the construction can be decomposed into separate parts, that is not the case. The only part that can be changed within this idiom is the number of years, and the remaining of it stays the same for re-use with other ages, granted that tense and subjects will change, but the lexical items involved in the construction remain the same. Contrariwise, one of the constructions that emerge in the data is UNEXPRESSED SUBJECT + TER-$_{\text{PRESENT INDICATIVE}}$ illustrated in (3). This construction is more open in that it is not processed as whole unit and can take any kind of subject. (Unexpressed subjects are given within parentheses in the English translation)

(2) Ø *<u>tenho</u> quarenta e um anos.*
'(<u>I) am</u> forty-one years old.'
(Inq. 34:194)

(3) *papai tem noventa e quatro anos...*

Ø tem treze filhos
'Dad is ninety-four years old...
(He) has thirteen children'

(C34: 200-201)

Since the usage-based model posits that language consists of the representation of psychological events, that is the speaker's experience with language, great importance is given to the way frequency affects the buildup of the speaker's linguistic system. In the next section, I discuss how frequency has been empirically shown to affect linguistic representation.

Givón claims that besides structural factors, there are pragmatic factors exerting functions on linguistic structures and their variation within a linguistic system (2001, p. 16). By taking this position in raising pragmatics to part of linguistic structure, or grammar, Givón opposes the traditional tripartite dogmas of structuralism: arbitrariness of linguistic sign, *langue* as an idealized system, and the strict distinction between synchrony and diachrony, which Givón sees as an extension of the linguistic system.

This idealization of the linguistic system is the basis of Linguistic studies that focus on virtual regularities of the system and neglect the particular use of speakers when using language in real time. Givón disagrees with this view in that by observing only the regularities, the mechanisms responsible for the constant reshaping of language and the linguistic system are ignored:

> (…) all functional-adaptive pressures that shape the synchronic – idealized – structure of language are exerted during actual performance. This is where language is acquired, and where grammar emerges and changes. This is where form adjusts itself – creatively and on the spur of the moment's opportunistic construal of context – to novel functions and extended meanings. This is also where slop, variation and indeterminacy are necessary

> ingredients of the actual mechanism that shapes and reshapes competence. (Givón, 2001, p. 6)

Givón has famously postulated that "today's morphology is yesterday's syntax" (1971, p. 413), which he later extended to include pragmatics by claiming that "today's syntax is yesterday's pragmatics" (2001). While these postulates have been shown to be true in various analyses, they are not inherently sufficient to explain linguistic variation, nor can it be implied that all grammatical changes are derived from pragmatics. Weinreich, Labov and Herzog have shown that not only pragmatic factors may play a role in linguistic change, but internal (linguistic) factors demonstrate an effect as well (1968). This suggests that the linguistic system is more adaptive that Givón proposes.

Givón condemns the analysis that attempts to "discover the pristine system hiding behind messy reality" (2001, p. 6). However, by establishing a form of linearity to explain linguistic change whereby grammar is subordinated to pragmatics, he appears to search for the same "pristine system" behind the chaotic reality that language is.

In short, it is worth mentioning that the view supported by Givón that grammatical structure is constantly being reshaped and remodeled is espoused in this study to the extent that linguistic production is considered the main source of the grammar of speakers. Thus, I propose that it is more fruitful to analyze linguistic change taking into account different systems and their subsystems, acknowledging that they are inter-related, but subordinated. So the sociolinguistic framework complements functionalism in that sense.

It is also important to note here that the use of the Variationist approach is completely aligned with the premises of Usage-based Linguistics in that grammar is identified in discourse through the observation of recurring patterns, and these patterns can be abstracted through the constraints and conditioning obtained in the statistical analyses. Thus, the pairing

of these two theoretical approaches complement each other in that Usage-based Linguistics provides the framework through which the patterns observed in the statistical analyses can be explained.

1.3.1 The importance of frequency

Usage-based linguistics postulates that linguistic items and structures are gradient and highly affected by input – e.g. frequency among others (Bybee, 2001, p. 20). In this sense frequency of input is crucial in establishing the relational connections within categories. As frequency of input increases, linguistic items become stronger and easier to access. Therefore, the storage of linguistic structures and lexical items is in part contingent on frequency effects. Storage is conceived not as a list of items but as a network of connections, which are strengthened between the items that share similar properties (Bybee, 1985).

The usage-based approach to linguistic analysis holds that the mental grammar of the speaker (his or her knowledge of language) is formed by the abstraction of symbolic units from situated instances of language use (Bybee, 2006). An important consequence of adopting the usage-based approach is that there is no principled distinction between knowledge of language and use of language (competence and performance in generative terms), since knowledge emerges from use. From this perspective, knowledge of language is knowledge of how language is used (Hopper, 1998).

Studying usage, then, especially frequency, can often tell us more about structure than attempting to study syntax as an autonomous entity (Bybee, 2002b). Bybee and Scheibman (1999), for instance, argue from phonetic data that high-frequency phrases containing *don't* (e.g., *I don't know*) do not adhere to the traditional constituency structure illustrated in (4). Having analyzed the distribution of the phonetic variants of *don't,* Bybee and Scheibman

show that phonetic reduction is most likely to occur when *don't* collocates with *I*, its most frequently co-occurring subject in discourse. In fact, when a full NP, or even another pronoun, is in subject position, *don't* reduction is highly unlikely. This pattern of phonetic coalescence, which is tied to particular co-occurrence patterns of English syntax, is an indication that, in some highly frequent collocations, the subject NP and the Aux form a tighter constituent than the Aux and the V do per traditional analysis.

(4) [NP] [(Aux) V], or [I] [don't know]

Krug (1998), furthermore, provides additional evidence from other pronoun-auxiliary contractions (*I'm, she's, they're*) that constituency in English does not always adhere to the structure outlined above, but rather reflects Halliday's notion of mood (2004, p. 72), as in the structure exemplified in (5), wherein the subject and auxiliary form a component unit. Krug, like Bybee and Scheibman (1999), demonstrates that the phonetics in his data are best accounted for with a frequency explanation. With this in mind, then, frequency of use seems to be driving constituent structure; otherwise, it could be predicted that English (and other languages) would always exhibit auxiliary contraction within, rather than across, traditional constituent structure as can be illustrated in (5) also discussed by Bybee and Scheibman in their analysis of 'I don't know.'

(5) [NP AUX] V

While it is true that frequency of co-occurrence can lead to semantically anomalous structures, it is also true that

> since contiguity in discourse is determined by pragmatic and semantic factors, items that occur together will be relevant to one another. In the usual case, then, this principle will lead to the commonly occurring constituent relations – preposition with NP, adjective with noun, auxiliary with verb, and so on. (Bybee & Scheibman, 1999, p. 593)

Both traditional and novel constituency structure result from on-line processing and chunking of frequent linear sequences and cognitive abilities such as blending. In other words, human language is not inherently organized in terms of logical syntactic structure nor is the human capacity for language necessarily endowed with innate structure-preserving processing abilities (Deutscher, 2000); instead, it is equipped with cognitive abilities to extract and create structure from the input. This position is well represented in Armstrong, Stokoe and Wilcox (1995) who argue that the human cognitive ability to identify a relationship pattern within a visual gesture provides a much more plausible scenario for the origins of syntax than those found in generative accounts of linguistic evolution (p. 184).

While we will never know exactly how early hominids communicated, we can assume, as in geology, the Principle of Uniformitarianism, which states that the processes observed in modern time are the same processes that operated in the past (Heine & Kuteva, 2007). Thus, it can be assumed that linguistic structure has from the beginning been the product of usage patterns and the cognitive abilities to extract those patterns as evidenced in the process of grammaticization. For example, a number of researchers (e.g., Bybee & Dahl, 1989; Bybee, Perkins, & Pagliuca, 1994; Heine & Kuteva, 2007; Hopper & Traugott, 2003 inter alia) have shown that similar grammatical items in distinct and unrelated languages can be traced back to recurrent usage patterns of specific lexical items. This *in situ* creation of grammar can be exemplified by further examining two syntactic categories: prepositions and auxiliaries.

One of the central claims in Cognitive Grammar, with respect to the usage-based model, is that usage affects grammatical representation in the mind. Furthermore, frequency correlates with entrenchment. Two main types of frequency effects have been described in

the literature: token frequency and type frequency. Each of these gives rise to entrenchment of different kinds of linguistic units.

Token frequency refers to the frequency with which specific instances are used in language. For instance, the semantically related nouns *falsehood* and *lie* have very different tokens frequencies. While *lie* is much more commonly used, *falsehood* is less frequent in use (Bybee, 2002c).

While token frequency gives rise to the entrenchment of instances, type frequency gives rise to the entrenchment of more abstract schemas. For instance, the words *copos* 'glasses', *gatos* 'cats', *cachorros* 'dogs' are all instances of the plural schema [NOUNs]. Other forms such as *talheres* 'silverware' and *mulheres* 'women' are instances of the plural schema [NOUN*es*]. As there are fewer usage events involving the second schema than there are of the first one, it is predicted that the former will be more likely to evoked by speakers because of its more generalized status, while the latter is less likely to be evoked in application to newer usage events. To cite an English example, let us consider the regularization of the past tense in English. The productive pattern in the language is done through the addition of *–ed* to the infinitival root of verbs (e.g., *work/worked*). However, a handful of verbs have retained an irregular continuum of patterns that range from vowel alternation (e.g., *drink/drank*) to complete suppletion (e.g., *go/went*). It has been shown, thus, that low-token-frequency verbs placed along the irregular continuum tend to regularize (e.g., *weep/weeped*) as opposed to high-token-frequency verbs, which tend to maintain their irregular pattern (e.g., *go/went*). Indeed, scholars have noted that irregular forms, which retain the morphology of an earlier stage of the language, tend to be the most frequent in a language (Bybee, 1985).

Bybee and Slobin (1982) provide empirical evidence for the view that frequency correlates with degree of entrenchment. They found that highly frequent irregular forms resist regularization, while irregular forms tend to become regularized over time. Bybee and Slobin compared irregular past tense forms of English verbs like *build – built*. They found that more frequently used irregular verbs like *lend* retain the irregular past tense form (*lent*). In contrast, less frequent forms like *blend* could alternate between the irregular form (*blent*) and the regular past form with the suffix *–ed*. Indeed, scholars have noted that irregular forms tend to be the most frequent in a language, and they tend to retain the morphology of an earlier stage of the language as well (Bybee, 1985).

Due to the non-reductive nature of the model, the predictability of an instance from a schema does not entail that the instance is not also stored in the grammar. Indeed, a unit with higher token frequency is more likely to be stored. For instance, the form *meninas* 'girls' is predictable from the lexical item *menina* 'girl' and the schema [NOUN*s*]. However, due to the high token frequency of the form *meninas* 'girls', this lexical item is likely to be highly entrenched, in addition to the form *menina* 'girl' and the plural schema [NOUN*s*] (Hay, 2001; Hay & Baayen, 2002, 2005).

Frequency of use, then, is crucial to understanding how our grammatical units originate. However, frequency of use is meaningless unless it is understood in terms of languages user's cognitive skills to infer and categorize. As Bybee (2006) explains, perceptual details, even mundane and predictable ones, are registered in memory (and Bybee, 1994; see also Langacker, 1987). If details occur repeatedly enough, these details are incorporated fully into the mental representation of a linguistic form. Therefore, speakers must have registered even early instances of a form.

In addition to cataloguing phonetic variation, linguistic context, and possible inferences about specific tokens, speakers also subconsciously track the type frequency of a construction, i.e. the number of different lexical items used with it. Bybee (1985, 1995) has shown that type frequency is a fundamental aspect of grammatical competence by documenting numerous examples in which type frequency correlates with a construction's productivity (Barddal, 2006; Barddal & Eythórsson, 2003; Barddal, Kristoffersen, & Sveen, 2011; Hay, 2001; Hay & Baayen, 2002, 2005). That is, if a construction or schema can be used with many different verbs, nouns, etc., the construction is more likely to be applied to novel forms (Bybee, 1985, 1995) or to be selected in language change (Bybee, 2003; Bybee & Thompson, 1997; Poplack, 2001). As Bybee and Thompson (1997) explain, human categorization and processing abilities are the cognitive basis for the structural fact that type frequency and productivity correlate:

1) as the number of lexical items used in a particular pattern increases, the less likely it is that the pattern itself will be associated with any one lexical item;

2) the more items that are used in a particular construction, the more general the features of the construction must be, thus allowing for even more (novel) members to be allowed in the construction; and,

3) the more items that are used in an open slot, the more often the construction will be used, strengthening the construction's representation and ensuring greater accessibility for future, potentially novel uses.

In addition, lexical connections serve to capture a construction's type frequency, that is, the number of different lexical items that fill a construction's open slot. The more lexical forms that are used with a particular construction, the greater productivity of that construction (Bybee, 1985). Thus, each time a construction is used with a new or different type, the connections that make up the construction are strengthened, making the

construction more "available… for the sanction of novel expressions" (Langacker, 2000, p. 26). In other words, as speakers map input to their mental representations, those constructions that are connected to different lexical types will have stronger representations and will not be associated with specific lexical expression. As a result, the constructions with higher type frequencies will be primed for subsequent use in conversation and extension to nonce forms. Thus, the type frequency of a pattern, and hence the degree of productivity of a pattern, is captured in the network of connections.

Thus, the human brain attend to details, such as phonetic variation, type and token frequency of use, linguistic context, possible inferences, etc. and store those details as part of a word's or construction's mental representation; if the details occur frequently enough, they will accrue, slowly altering the mental representation over time and giving rise to new structures.

The linguistic unit in a usage-based model, then, encompasses a much wider range of linguistic expressions than a dictionary conception of a lexicon does. As a result, no limit is placed on what can or cannot exist in the lexicon: morphologically complex words, common phrases, idioms, chunks, collocations, lists, clauses, constructions, even entire passages can be redundantly stored as abstract units (Goldberg, 2006; Van Lancker, Kreiman, & Bolinger, 1988; Wray, 2000). However, the usage-based lexicon is not an unordered list of linguistic units. Instead, units are organized within a network of phonologically and/or semantically related forms which are connected via lexical links (Bybee, 1985). These lexical connections identify recurring form and/or meaning patterns across linguistic units in the lexicon. As a result, the connections support the emergence of structure.

Frequency has been noted to affect the representation of linguistic forms, and/or constructions, in several distinct ways, such that it plays a key role in language change (c.f. Bybee & Thompson, 1997 for a survey of some of these effects). An important token frequency effect, also known as the conserving effect (Bybee, 2006, p. 715), concerns the entrenchment of a structure rendering it more resistant to restructuring based on productive patterns.

The notion of a conserving effect of frequency is extremely important to understanding the variation at hand. Even though it will be seen that both 1sg and 2sg subjects have become increasingly expressed more frequently over the years to the point that it can be argued that this is now the canonical pattern, 3sg subjects have retained their older pattern, that is lack of expression, and it is contended here that the conserving effect is partly accountable for this. Moreover, 3sg subjects need to be viewed as a heterogeneous category, if it is in fact a category at all, thus the possibilities of patterns forming between these subjects are sparser (R. M. W. Dixon, 2009). This category not only encompasses animate, but also encompasses inanimate pronominal referents. The former can be realized with *ele* 'he' and *ela* 'she', while the latter can be realized with these two pronouns as well as with an array of others that have other functions, such as the demonstratives *isso*, *isto* and *aquilo*, neuter 'this' and 'that'.

1.3.2 Constructions

Constructional approaches to linguistic description are defined by two key properties. Scholars working with constructional approaches agree that the units of grammar are symbolic, that is to say they are conventionalized relationships between forms and meaning. They also agree that there is no real distinction between "core" phenomena central to

grammar and "peripheral" phenomena which are not so central (Chomsky, 1965). These two properties make constructional approaches particularly relevant to the description of languages and the patterns that emerge from usage.

There are different conceptions of the constructional agenda, and I will try to describe some of them in this section. Some constructional approaches are to be found at the relatively non-formal and functionalist end of linguistic theorizing; others are highly formalized and do not have a great deal to say about functional pressures in language. Some constructional approaches restrict their assumptions to a willingness to admit non-compositionality to the ontology of their grammatical theories; others assume that language is usage-based, and that non-compositionality is not the only basis for taking a constructional approach.

However, these different background assumptions of scholars working with constructional approaches, the different views of what should be in a constructional theory of grammar, do not affect the utility of constructional approaches to language-particular description, as it is the case with subject expression. Once it is agreed that grammar is symbolic, the issue becomes identifying the symbols of the grammar of the language being investigated. Thus, this conception makes the constructional approach particularly apt for language-specific description.

To explore construction grammars, I will start by looking at some of the central claims and how they pertain to the issue being addressed in this dissertation. First, grammar is symbolic, in that words are relationships between forms and meaning (Bergan & Chang, 2005; Bybee, 2010; Bybee & Cacoullos, 2009; Croft, 2001; Fillmore, Kay, & O'Connor, 1988; Goldberg, 2006; Goldberg, Casenhiser, & Sethuraman, 2004). A noun, for instance, has a phonological shape, syntax, a sense, and a referent. The first two are part of the word's

form and the last two are part of its meaning. In some theories of construction grammar, morphemes are likewise constructions (Croft, 2001; Goldberg, 1995, 2006). According to Croft and Cruse (2004), a clause or a sentence, or the subject, all instantiate form-meaning pairings which involve conventional units that are larger than individual words.

The second major claim follows from the observation that there are limits to compositionality (Hay, 2001; Hay & Baayen, 2002, 2005). In their seminal paper, Fillmore, Kay and O'Connor (1988) explored idiomaticity, demonstrating that there is partial regularity and partial compositionality. Nevertheless, there is also an element to the meanings which is not predictable, and which suggests that they are not simply compositional. It is the status of not belonging to one or the other provides evidence for a constructional approach. As Nunberg, Wason and Sag (1994) point out, it is not the case that idioms are fixed expressions with fixed meanings.

Given that idioms exist, and given that they have their own meanings, it follows that there are constructions, that is, units of grammar which are larger than words[5], which are meaningful, and whose meaning is not regularly predictable from their parts. This observation is the second major motivation for construction grammars.

Constructional approaches to grammar are particularly relevant to language-particular research, largely because of the research agenda and underlying assumptions of constructional approaches such as Goldberg (1995, 2006), Bybee (2010), and Croft (2001), which tend to focus on important phenomena within individual languages. This is because

[5] This is not imply that there are not units smaller than words, on the contrary, by units here we mean items that produced through one processing strategy.

constructional approaches assume that languages are structured out of conventionalized form-meaning pairings at all levels of grammatical description.

In this dissertation, constructions are assumed to play an important role in the shaping of pronominal expression in BP. Any highly frequent combination of subject, verb and tense is considered a construction in this work. By highly frequent, I mean constructions that meet a threshold of frequency in the corpus, which in this study corresponds to one percent of the data for each person. Although, this percentage threshold is arbitrary, it has been suggested by others as a starting point for this type of investigation (Goddard, 2005). Thus, it is argued that the choice of pronominal subjects versus unexpressed subjects is largely due to the patterning, or constructions, of subjects, predicates, and tenses in the data.

1.4 Overview of Variationist Theory

In this research I adopt the variationist framework (e.g., Labov, 1969, 1972a, 2001; Poplack, 1993; Poplack & Tagliamonte, 2001; Sankoff, 1988a, 1988b; Tagliamonte, 2006, inter alia), which seeks to discover patterns of use by employing quantitative techniques to determine the effect of contextual factors on the choice of a form. To analyze and understand the mechanisms involved in the variation between expressed and unexpressed subjects in BP, I invoke the theoretical approach and tools of Variationist Theory. In this section, I will establish the tenets of this approach that inform this study and outline the methodological principles involved in this theory.

Sapir (1949) asserted that the phenomenon of language variation induces changes in the language, in other words, if there two or more forms are in competition for a similar linguistic function, one will eventually overcome the other in being the preferred choice for such function, thus creating a change in the language. This tenet was captured in the seminal

work by Weinreich, Labov, and Herzog (1968), who postulate that the primary object of linguistic investigation is the speech production of speakers of a particular linguistic community. In short, it is necessary to investigate language within its community, accounting for the interaction between linguistic forms and social contexts.

The linguistic community is seen as a group of people who share overall patterns of use, but not a group of speakers who speak in the same way (Labov, 1972a, 1994, 2001). Despite the fact that these speakers share the same language variety, and that their speech exhibits the linguistic resources available to them, their grammar may still demonstrate great levels of variation, which represents a systemic heterogeneity in that while language, or grammar, shows variation across and within speakers, this variation is systematic and can be described.

It is in this context that variationist linguistics makes another important contribution: it shows that such heterogeneity in terms of linguistic forms used by speakers is not random or chaotic. Rather, it is part of an inner system that can be identified and described through empirical research. Thus, a theoretical approach that aims at dealing with language variation and change must be able to cope with an ordered heterogeneity, which is a fundamental characteristic of language (Labov, 1994):

> The key to a rational conception of language change – indeed, of language itself – is the possibility of describing orderly differentiation in a language serving a community. We will argue that nativelike command of heterogeneous structures is not a matter of multidialectalism or 'mere' performance, but is part of unilingual linguistic competence. One of the corollaries of our approach is that in a language serving a complex (i.e. real) community, it is **absence** of structured heterogeneity that would be dysfunctional (Weinreich, et al., 1968, p. 101).

Labov then provided the methods and theoretical tools necessary to establish this kind of analysis. According to the author, a linguistic system does not consist only of rules or categorical elements, that is, rules that are always applied and the categorical elements that

are always realized in a particular manner, but it also contains elements that are in variation. The latter are called linguistic variables, and they may correspond to two or more elements.

The linguistic variable is defined, thus, as distinct possibilities of expressing the same concept, in the same contexts, with the same truth value. To put in another way, the linguistic variable can be expressed as different ways of saying the same thing. They are, therefore, similar in their reference, though they may differ in their social value and/or in the linguistic environments in which they occur (Labov, 1971, 1994).

The alternation between the realization or lack thereof of pronominal subjects in BP is a classic example of a linguistic variable. From the variationist perspective, this variation is systematic and non-random inasmuch as it is conditioned by both internal (linguistic) (c.f. Duarte, 1993; Lira, 1982; Paredes Silva, 1993 to name a few) and external (non-linguistic, in particular, social) factors (c.f. Monteiro, 1990, 1994b; Rollemberg, Andrade, Lopes, & Matos, 1991 inter allia). While it is acknowledged that external factors play a crucial role in the conditioning and realization of any linguistic variable, in the present study, however, only the internal, i.e., the linguistic factors will be analyzed. This is so because of the nature of the linguistic variables that are being analyzed in that they are examined taking into account the frequency of the verbs with which subjects most occur and the effects of constructions of subjects, verbs and TAMs have on each individual person and their rates of subject expression. Thus, the main objective of this work is to look at the effects of frequency and constructions in the conditioning of linguistic factors and how they shape the way pronominal expression is borne out in BP.

Labov's model of language variation and change presupposes that variation and change are intrinsically related. The processes of change that one identifies within a linguistic

community can be updated and retrieved in different moments in time by examining the speech of different speakers. However, the presence of variation does not predictably suggest change (Weinreich, et al., 1968). And this is one of the findings of this research. While there seems to be an apparent change in progress toward expression in BP (as has been suggested by Castilho, 1987; Duarte, 1993, 2000, 2003; Tarallo, 1993, inter alia), what is found is that each person is patterning differently with some high frequency verbs, and such patterning demonstrates a different behavior with relation to pronominal expression different from that observed in the remaining of the data.

Linguistic change is motivated both linguistically and socially. So, linguistic change or variation must rely on both external and internal factors to explain the forms that emerge in discourse. However, the study of linguistic change is against the view that the grammar of speakers is a finished product and is therefore not susceptible to further changes within its structure (cf. Newmeyer, 1998; Newmeyer, 2003 for discussion). Thus, one should not describe grammar as a fixed system, but rather, as an emergent one (Hopper, 1987, 1988, 1998).

Furthermore, Labov notes that language change evolves from a disruption of the relationship between form and meaning in that speakers affected by the change do not purport the same meaning as those not affected by the change (i.e. older speakers or speakers from other communities) (1994, p. 9).

With this notion of language change in mind, numerous scholars (Barbosa, et al., 2005; Castilho, 1987; Duarte, 1993, 2000, 2003; Galves, 2000; M. Modesto, 2000a; Monteiro, 1990, 1994b; Negrão & Müller, 1996; Raposo, 1998; Tarallo, 1993) have proposed that BP is undergoing a change in its verbal paradigm which is also leading the

change from a formerly pro-drop to non-pro-drop language. Thus, the study of language change in progress, proposed in this study, has also the aim of contributing to the better understanding of how these forms evolve. While this is study is synchronic in its nature, I will attempt to offer explanations to the possible change this phenomenon is going through.

The principle of Uniformitarianism posits that changes in the past can be explained by corollaries found in changes in the present (Christy, 1983). If we indeed accept that changes in the past are governed by similar principles observed in the present, than it follows that

> The same mechanisms which operated to produce the large-scale changes of the past may be observed operating in the current changes taking places around us (Labov, 1994, p. 161).
>
> Knowledge of processes that operated in the past can be inferred by observing ongoing processes in the present (Christy, 1983 apud Labov, 1994:21).
>
> The factors that produced change in human speech five thousand or ten thousand years ago cannot have been essentially different from those that are now operating to transform living languages (Labov, 1994, p. 22).

Thus, one can use the past to understand the present as one can use the present to understand the past. Examining ongoing linguistic changes provides us with the tools to understand the mechanisms through which they got to where they are synchronically.

It must also be noted that it is through data, that is to say, language produced in real circumstances that reveal the true nature of the grammatical system of a given language. Through language obtained in real time, it can also be observed the pathways to change, provided that these changes have some form of social motivation.

2 SUBJECT EXPRESSION IN BRAZILIAN PORTUGUESE

2.1 Subjects in Brazilian Portuguese

The concept of subject proposed here is that it is a grammatical relation that is the normal expression of the grammatical functions A, or the more agentive role in a two-argument clause, and S, or the single argument in a one-argument clause (B. Comrie, 1981; R.M.W. Dixon, 1979; Du Bois, 1987, 2003; Du Bois, Kumpf, & Ashby, 2003). In Brazilian Portuguese, as in many other languages, there are a variety of coding features that distinguish subjects from other grammatical functions such as obliques. Namely, these coding features include the nominative, preverbal position and verbal agreement and in (6) as opposed to the accusative case as in (7), and elision as in (8) (Ilari, Franchi, Moura Neves, & Possenti, 1996; Monteiro, 1994b; Perini, 2002).

(6) *Ele* *já* *está* *matriculado* *no* *Batista.*
He *already* *is enrolled* *at* *Batista.*
NOM.SG *3SG*
'He is already enrolled at Batista.'

(C7: 606)

(7) *Ele* *vai* *fazê-lo.*
He is going to do - it
NOM.SG 3SG ACCU.SG
'He is going to do it.'

(C116: 483)

(8) *...Ø tão percebendo agora, né?*
'… (you) are noticing now, aren't you?'

(L53: 312)

BP, when compared to other languages, is considered to be part of a group of languages that allow for *pro-drop*, or the elision of arguments, as opposed to other languages that do not demonstrate such patterning. Even though BP and European Portuguese (EP) are

mutually intelligible dialects, they show very distinct rates of expression, with the former demonstrating a much higher rate than the latter (Barbosa, et al., 2005).

Galves (2000) and Kato (1999) have provided generative accounts of the phenomenon, while Tarallo (1993), Duarte (1993), Lira (1982), and Paredes Silva (1993) have examined subject expression under the framework of functional and variationist linguistics.

In the literature on subject expression in BP, there has been a major claim establishing a corollary between rich verbal morphology and omission of subjects and weak verbal morphology and expression (Duarte, 1993; Mary Aizawa Kato, 1999; M. Modesto, 2000a).

This hypothesis seems particularly applicable to BP because it is undergoing a simplification of its verbal morphology. Hence, from a paradigm of six forms (*canto, cantas, canta, cantamos, cantais, cantam*, 'I sing, you sing, he/she sings, we sing, you sing, they sing') the system has reduced to four forms with the substitution of *você* for *tu* and *vocês* for *vós* (*canto, canta, cantamos, cantam*, 'I sing, you/ he/she sings, we sing, you/ they sing'), and further to three forms with the new substitution of *a gente* for *nós* (*canto, canta, cantam*, 'I sing, you/he/she/we sings, you/they sing') (A. T. T. Modesto, 2006; Travis & Silveira, 2009).

Duarte (1993) describes the impact of this impoverishment of verbal morphology on variable subject expression. Based on plays written by Brazilian playwrights from the nineteenth and twentieth centuries, she observes a decline in the rates of subject expression over time (cf. section 2.3.1 for more detailed discussion) (Duarte, 1993, p. 117). Her results suggest that the functional category of agreement no longer behaves as a predictor of subject

expression in BP. Instead, it evokes a correlation that less agreement morphology equates to more subject expression. Thus, BP is becoming a language where zero arguments are constrained to certain limited environments. According to Duarte (1993), unexpressed subjects can still be found in the following contexts:

- with first person singular subjects "em orações independentes com verbos simples no presente ou passado, quase sempre precedidos por uma negação, ou com uma locução verbal[6]" (Duarte, 1993, p. 119) as in

  (9) ***Não posso mais ficar*** *aqui a tarde toda, não, tirei quatro notas vermelhas,* ***preciso dar*** *um jeito na minha vida.*
  '(I) can't stay here all afternoon, no, (I) got four bad grades, (I) need to do something about my life.'

- with first person singular in subordinate clauses

  (10) *Eu não sei se* ***vou conseguir*** *numa sessão só.*
  'I don't know if (I) will manage it in one session only.'

- with second person singular in interrogative sentences

  (11) *já se esqueceu?*
  'Have (you) forgotten it already?'

  (12) *falou com ele?*
  'Have (you) spoken with him?'

Kato (1996) arrived at similar results in support of the correlation between impoverished morphology and the rate of subject expression. She examined data from the project NURC (Norma Urbana Culta) and found that only 19% of first person singular subjects were unexpressed. These findings support Duarte's (1993) study and further show that unexpressed first person singular subjects can also occur in coordinated clauses, with

[6] "in main clauses with the main verb in the present or past preceded by a negation marker or in sequences of auxiliary and verb."

unaccusative verbs (e.g., *chegar* 'to arrive', *entrar* 'to enter', and *partir* 'to arrive') (cf. Clements, 2006 for examples), and with a verb whose direct object position is already filled.

Negrão and Muller (1996) have also attempted to explain the variation between pronominal and unexpressed subjects in BP. They begin by saying

> se o enfraquecimento da flexão é a causa do preenchimento progressivo da posição do sujeito, esperaríamos que o aumento de preenchimentos se desse especialmente naquelas pessoas para as quais a morfologia verbal não é mais capaz de identificar o sujeito (2ª e 3ª pessoas). Esperaríamos, também, uma maior proporção de preenchimento para os casos em que há ausência de "concordância", ou seja, em que a pessoa do verbo não é a mesma que a do sujeito.[7]

They hypothesize then that

> estaria havendo uma especialização no sistema pronominal do PB segundo o tipo de denotação semântica que se deseja expressar. O pronome *ele* e a forma possessiva *dele* são usados para expressar sintagmas nominais referenciais. A categoria vazia não arbitrária na posição de sujeito e a forma possessiva *seu* seriam usadas para expressar uma relação anafórica entre estes sintagmas nominais e seus antecedentes.[8]

[7] "if the impoverishment of the verbal morphology is the cause of the progressive increase in expression, it would be expected that this increase would take place within those persons where the verbal morphology can no longer identify the subject (2nd and 3rd). It would be expected as well that there would be an increase in the rates of expression for those cases where concord is absent, that is, where there is a mismatch between the person marked in the verb and the referent."

[8] "there seems to be a specialization in the pronominal system of BP according to the semantics the speaker wishes to express. The pronoun *he* and the possessive form *his* are used to express referential NPs. Unexpressed subjects and the possessive form *your* are used anaphoricaly between Nps and their antecedents.

Thus, it is necessary to observe "os mecanismos de identificação do conteúdo referencial das formas pronominais de uma determinada língua[9]" (Negrão & Müller, 1996, p. 148), meaning that the argument needs to move away from syntax and be explored at a semantic level and discourse level.

These explanations rely on the intution of the linguist analysing the phenomenon rather than on real-time data produced by speakers. Thus, these analyses invoke factors that are more general and formal in nature, for example, impoverished agreement.

While it appears that impoverishement of agreement correlates with the increase in pronominal expression in BP, they merely correlation. Studies have only showed that there has been na increase in the rates of expression. Without really making a Strong connection between the two. Moreover, the impoverishment in the agrément system has been mostly concentrated with 3sg agreement, where the rates of expression have changed very little, which is again na indication of the coincidental correlation between the two. In short, the impoverishement of agreement and the increase in pronominal subjects in BP appear to be concomitant changes stemming from diferente changes in the language (e.g., the inclusion of *você* 'you' and *a gente* 'we' in the pronominal system).

2.2 Subject realization in Brazilian Portuguese

In BP, the head of an NP is typically expressed by a common noun, a proper noun, or a pronoun. Usually only common nouns accept modification (R. M. W. Dixon, 2009). Pronouns occur alone and proper nouns can be preceded by a definite article, as in (13) below.

[9] "the mechanisms in which pronominal forms establish referential content within the language."

(13) *A China usa irrigação à larga.*
'China uses irrigation extensively.'

(I10: 502)

Moreover, there are many Portuguese clauses that occur categorically with an unexpressed subjects[10], as in (15), where the information about the subject referent cannot be retrieved from the verb inflection or from the context as opposed to (14) where the referent can be retrieved from previous context and discourse.

(14) *Ela estava muito gorda… tava desproporcional para a idade dela, acredito que ela comeu muito doce quando era pequena.*
'She was very overweight... (she) was disproportional for her age, (I) believe that she ate a lot of sweets when she was a child.'

(I9: 831)

(15) *Ultimamente só Ø chove aqui*
'Lately (it) only rains around here.'

(C30: 732)

The characteristics of the subject NP in BP briefly outlined above allow me to divide the Portuguese subjects into three main categories of realization in order to discuss the variation being investigated here: nominal subjects, pronominal subjects, and unexpressed subjects. The distribution of these three kinds of subjects in the data under study here can be seen in Table 1 below.

[10] There are sentences in Portuguese which are subjectless, the verbs involved are (a) nature verbs, e.g. *chover* 'to rain', *trovejar* 'to thunder', (b) the verb *haver* 'there to be' and *ter* 'to have' with a similar sense, and (c) the verbs *haver* 'there to be', *ser* 'to be', and *fazer* 'to do' expressing time.

Table 1. *Distribution of subjects across three realization types in BP*

		N
1sg		
	Pronominal	2262
	Unexpressed	1185
2sg		
	Pronominal	911
	Unexpressed	778
3sg[11]		
	Pronominal	
	Animate	1400
	Inanimate	485
	Unexpressed	
	Animate	1530
	Inanimate	1767
	Lexical	2482
Exclusions		6010
	Total	**18,810**[12]

As can be seen from Table 1 above, pronominal subjects in BP are quite frequent, accounting for 66% of 1sg subjects, 54% of 2sg subjects, and 48% of human 3sg subjects. There are many types of pronouns, which can function as subjects: indefinite pronouns, relative pronouns, demonstrative pronouns, and interrogative pronouns, as well as personal pronouns. In this study, the indefinite (e.g., *algum* 'some') and demonstrative pronouns (e.g., *esse/a* 'this') were classified under nominal subjects following Lira (1982, p. 76). The relative (e.g, *que* 'that' and *quem* 'who/whom') and interrogative pronouns (e.g., *qual* 'what', *como* 'how') were excluded since they involve many complex transformations which deserve special attention by themselves and are thus beyond the scope of this study

[11] These subjects include the pronouns *ele* 'he' and *ela* 'she', which pattern similarly.

[12] It must be noted here that this number does not imply that all these tokens were included in the statistical analyses. This figure illustrates a raw count of subject realizations in the dataset culled from the corpus. In Chapter 3, I discuss a series of exclusions to which this data was submitted, leaving us with a total of 8,066 tokens to be analyzed statistically in this study.

(Cameron, 1995; Otheguy, et al., 2007). Under the head of pronominal subjects I included the personal pronouns *eu* 'I', *você* 'you' and *tu* 'you', *ele/a* 'he/she/it', *senhor/a* 'sir/madam', which have been fully grammaticized as pronouns in BP and they form the focus of this study.

2.3 Previous accounts of subject expression in Brazilian Portuguese

When faced with the question of why some languages allow unexpressed, or "null"[13] subjects, but others do not, most people tend to hypothesize that, in languages like Spanish and Portuguese, the information about person and number is directly recoverable from the verbal inflection, which makes an expressed subject unnecessary. In languages like English, on the other hand, an overt pronoun must occupy the subject position in order to disambiguate the sentence. This intuition was formalized by Taraldsen (2006)[14]. Since then, languages in which the verbal inflection determines, or recovers, the content (or the reference) of the subject have been called "rich" agreement languages. The relation between "rich" agreement and null subjects has been assumed in some form or another by many

[13] This term is used here to maintain consistency with the terminology employed by the scholars in the studies that I review. Other terms are unexpressed subject and empty subject. Throughout this work the term unexpressed subject will be used for the term null subject implies an affiliation with a theoretical perspective that is not followed here; moreover, the term empty subject seems to induce the reading that the category is non-existent when, in fact, this "emptiness" of subject is a result of the continuity of the form as the topic of discourse. As has been argued by Givón and others, the longer a form maintains a topical status, i.e., is retained in the forefront of the conversation, the more attenuated linguistic forms will be employed to express it (Chafe, 1994; Du Bois, 1987; Givón, 1983a, 1983b). It must also be noted that this statement is simply not true, in fact, South-east Asian languages are overwhelmingly isolating and omission of Pro is their common feature (Goddard, 2005).

[14] Taraldsen's generalization: *pro* is licensed if agreement is sufficiently rich to recover its features (p. 630).

linguists (Barbosa, 1995 for Portuguese; Chomsky, 1982; Jaeggli & Safir, 1989; Kenstowicz, 1989 for Arabic; Platzack, 1987 for Scandinavian languages; Rizzi, 1986 for Italian; Turan, 1995 for Turkish).

The validity of the claim that "rich" agreement is involved in determining if an argument may have no phonetic realization in a given language is supported by data from languages like Pashto (Huang, 1984, pp. 535-536). In sentences in the present, Pashto uses a nominative-accusative system: the verb agrees with the subject in both transitive and intransitive sentences. In past tense sentences, however, the verb agrees with the subject if intransitive, but with the object if transitive:

(16) a. Jan ra-z-i.
John DIR-come-3rd m. sg.
'John comes.'
b. zχ mana xwr-χm.
I apple eat-1st m. sg.
'I eat the apple.'

Other languages also provide support for the "rich" agreement idea when compared with Romance, where the verb agrees only with the subject and only subjects may be left unexpressed. In Swahili, for instance, the verb agrees with the subject and the object, and both these arguments may drop. In Basque, the verb agrees with every argument, and everything may drop.

Despite all the evidence supporting the relation between "rich" agreement and null arguments, such an idea is not devoid of problems. As noted in the literature, the property that makes agreement "rich" is difficult to pin down. Most researchers use the term "rich" to mean enough morphology to provide non-ambiguous information on the person and number

of the subject. However, this raises the question of how rich the inflection must be, or how rich is rich enough, to license null arguments.

To cite a well-known example, agreement seems to be rather rich in German, yet, null referential subjects are not permitted, and only non-referential null subjects are allowed. This fact has been captured by assuming that, in German, an empty category is licensed in subject position but not identified with referential features, so they are only possible when pleonastic (Harbert, 2007; Rizzi, 1986). To reinforce this argument it should be mentioned that there are many South-east Asian languages that allow for unexpressed subjects but do not have any agreement (Goddard, 2005; Ono & Thompson, 1997). Thus, it is argued that the relationship between agreement and unexpression must be questioned.

Based on the parameter system discussed in Huang (1984), according to which natural languages can be either discourse-oriented or sentence-oriented, Negrão and Viotti (2000) propose that BP should be considered a language of the first type. This same idea was proposed, in different forms, by Galves (2000). Discourse-oriented languages are described by Negrão and Viotti in the following way:

> A discourse-oriented language makes visible in overt syntax some relations that other languages only express in Logical Form. Among such relations are the informational function of certain constituents (such as discourse topic and focus), and the scope of quantifier phrases.

In discourse-oriented languages, the basic predicative relation is not one that is established between the subject and the predicate within the sentence, but one that is established between the whole sentence and a constituent that is outside. According to Huang (1984), one of the basic differences between discourse-oriented languages and sentence-oriented languages is that, in the latter, the most prominent element in the sentence is the

subject, whereas in discourse-oriented languages, the most prominent element in the sentence is the topic (Negrão & Viotti, 2000, p. 106).

In this way, in discourse-oriented languages, it can be seen that what licenses null subjects is not necessarily how rich the agreement system of the language is, but indeed the continuity of the topic in discourse. In other words, topics that are often reiterated in conversation become more accessible to both the speaker and the hearer, being thus more easily retrievable in discourse. Such forbearance of retrievability allows speakers to "drop" the subject in subsequent clauses even in contexts of ambiguity. I will now turn to some of the generative applications to explaining the phenomenon of subject expression in BP.

2.3.1 Non-functional accounts

As was discussed in the previous section, there is a lot of discussion concerning the fact that some languages possess a well-defined morphological system to mark person, number, or gender among other properties, whereas there are other languages that show no such markings, being considered to have a poor paradigm. In the case of Romance languages, it is believed that the verbal paradigm present in these languages is a rich one, and it is one of the major arguments for the nature of unexpressed pronominal subjects.

European Portuguese (EP) presents a rich morphological verbal paradigm and it shows a high frequency of unexpressed subjects (Barbosa, et al., 2005). BP, on the other hand, is undergoing a series of changes in its verbal paradigm that is resulting in a restructuring of its pronominal system. This reduction or restructuring is considered by some scholars as the major force at work in the increase of frequency of pronominal subjects in BP (Duarte, 1993, 2000, 2003; Tarallo, 1993). This increase, these scholars suggest, is leading BP toward becoming a language that does not fit into the *pro-drop* parameter. Other scholars,

on the other hand, contend that these changes are contributing to the emergence of a new form of *pro-drop* than the one seen in other Romance languages (Galves, 2000; Mary Aizawa Kato, 1999, 2000; M. Modesto, 2000a, 2000b).

Duarte (1993, 2000, 2003) has shown that this paradigm is changing in BP, which seems to be moving toward an obligatory subject language. In her study of subject expression in plays from the 1800's to the 1990's, she demonstrates the speed with which the language is changing. For example, unexpressed 1sg subjects go from a rate of occurrence of over 80% in 1882 to less than 20% in 1992. Similar changes can be seen for 2sg and 2pl as well as for 1pl (cf. Zilles, 2005 for discussion), and a steady increase in rates of expression for 3sg and 3pl are also found, but they do not reach the same status of 1sg or 2sg. In short, Duarte concurs with the previous analyses in that the increase in the rates of subject expression is a result of the change in the verbal paradigm. She contends that BP is losing its null-subject parameter and this is the result of the weakening of the verbal paradigm and the changes in the pronominal system. Thus, she concludes that the realization of zero subjects in BP is no longer a rule, but a variable that favors the overt subject as she puts it (Duarte, 1993, p. 141):

> Os resultados a que a análise variacionista nos permitiu chegar revelam que o português brasileiro perdeu a propriedade que caracteriza as línguas de sujeito nulo do grupo pro-drop por força do enfraquecimento da flexão, responsável pela identificação da categoria vazia em línguas que apresentam uma morfologia "rica" para tal processo, confirmando a hipótese de Roberts[15].

[15] "The results produced by the variationist analysis allowed us to claim that Brazilian Portuguese lost the property that characterizes null-subject languages. This loss is due to weakening of verbal inflection, which in turn is responsible for the identification of the empty category present in morphologically rich languages, confirming the hypotheses posed by Roberts."

Duarte further notes that despite the fact that all persons have exhibited a noticeable increase in their frequencies of expressed subjects, 3sg seems to be impervious to the change in the paradigm, in that these subjects continue to show high rates of expression (2000, p. 116). Her findings suggest that 3sg pronouns seem to adhere to the constraints established by traditional analyses.

Kato (1999) suggests that the unexpressed subject nature of BP is a result of its rich inflectional paradigm coupled with a rich pronominal system. According to the author, in languages that allow subjects to be absent, pronouns are coupled with agreement markers (inflectional suffixes in the case of BP) on the verb to establish co-referentiality. However, she suggests that impoverishment of agreement brings about the emergence of weaker or unstressed pronouns, which appear to function as clitics. These in turn tend to be expressed more frequently.

Table 2. Verbal agreement in Brazilian Portuguese

		Old system		New system	
1sg	*eu*	*Falo*		*eu*	*falo*
2sg	*tu*	*falas*	}	*você*	*fala*
3sg	*ele/a*	*Fala*	}	*ele/a*	*fala*
1pl	*nós*	*falamos*		*a gente*	*fala*
				nós	*falamos*
2pl	*vós*	*falais*	}	*vocês*	*falam*
3pl	*eles/as*	*falam*	}	*eles/as*	*falam*

Thus, the author argues that the change is a result of the resetting of the agreement system of BP, which moved from a set of six person inflectional suffixes to a set of 4 (see **Table 2** above), thereby inducing different persons of discourse into sharing the same inflectional marking, namely zero agreement for both second person singular (2sg) and third person singular (3sg) due to loss of /s/. Observe examples (17) and (18) below. In (17) it can be seen that the inflected verb (underlined) is realized in the same form as the verb in (18) in spite of the subject of (17) being a 2sg pronoun, and the subject in (18) is a 3sg pronoun.

(17) *Você vai se moldando né?*[16]
'You keep shaping yourself up?'

(C30: 206)

(18) *que ele num vai pegar o cara "EI PÁRA AÍ" ele vai pára?*
'that he won't grab the guy "hey stop there" and the guy will stop?'

(C11: 175)

In short, scholars who follow a generative approach account for the phenomenon of subject expression in BP in relation to the changes in the verbal paradigm. Although it is accurate to assume that changes in the morphology of a language may explain changes in the pronominal system, as the one contended here (Milroy, 1992; Naro & Scherre, 1991), this cannot be the only explanation for the increase in the occurrence of expressed pronouns in BP because it fails to explain how the change may take place. Plausible as this explanation may be, it does not explain why ambiguous verb forms nevertheless occur with unexpressed subjects. Hence the advantage of the Variationist approach, which allows us to observe how one factor group such as this interacts with other factor groups in conditioning the realization of the variable. Moreover, I will argue throughout this work, the frequency of certain verbs, and the way they pattern with expressed or unexpressed subjects condition the high rates of subject expression in BP.

2.3.2 *Functionalist and Variationist accounts*

As opposed to Generative explanations, variationist and functional accounts do not rely on the linguist's intuition[17] to elucidate linguistic phenomena and do not consider isolated

[16] There is also a historical explanation for the fact that *você* 'you' in BP, currently a 2sg pronoun, takes 3sg verbal agreement form. This pronoun is derived from a third person expression, literally meaning "Your Mercy" (Faraco, 1996).

examples out of context. Rather, linguists use large corpora, preferably of spoken language, to be able to examine the linguistic phenomena firsthand as it is produced by speakers. In functional linguistics, the emphasis is given to the purposes of using particular structures, rather than the mere structural characteristics of these structures.

A great number of scholars have examined the variability between expressed and unexpressed subjects in Romance languages. In this review I will focus primarily on the literature concerning the findings in Brazilian Portuguese drawing from related findings in Spanish as well. The research in these two languages points to a set of factors that appear to have an effect on the realization of expressed subjects (Bentivoglio, 1987; Cameron, 1992; Cavalcante, 2001; S. Cunha, 2003; Lira, 1982; Monteiro, 1994b; Otheguy, et al., 2007; Paredes Silva, 1993, 2003; Silva-Corvalán, 1982; Silveira, 2008; Travis, 2007). These findings suggest that there are discourse and functional factors that condition the distribution of subjects – rather than the traditional analysis claim of "rich" agreement. For the remainder of this section I will discuss some of the findings in this research that attempts to explain the nature of subject expression.

A number of different factors have been noted to affect subject expression in BP. Contrary to what has been found for the social factors, in that age, gender, and register are strong predictors of the realization of expressed subjects, there is not general agreement on what factors clearly affect this variability. Firstly, this is the result of different scholars using different types of data. To illustrate this point, consider the studies conducted by Lira (1982) and Paredes Silva (1993) who obtained very different results. The former analyzed spoken

[17] While Duarte (1993 and 2003) looked at data, she still posited that language use does not play a role in the way language changes and/or is manifested in the speakers produce it.

Portuguese from Rio, whereas the latter examined written Portuguese from the same dialect. This difference in the data examined provoked different factor groups to emerge as conditioning the realization of expressed subjects. The only agreement between the authors lies in the realization that different persons have different distribution between expressed and unexpressed subjects (c.f. also Otheguy, et al., 2007)[18].

Lira (1982) examined the speech of speakers from Rio de Janeiro and showed that the following factor groups favor expressed subjects:

a) Person;

b) when the subject of the preceding clause is a distinct one;

c) Relative and adverbial clauses;

d) Emphasis;

Paredes Silva (1993) observed informal letters from speakers from Rio de Janeiro. So, from the outset there is a difference in the population investigated by Lira and that investigated by Paredes Silva. Whereas Lira's data were drawn from speakers of a lower level of education, through sociolinguistic interviews, Paredes Silva's data was bound to be more formal because its written nature allowed speakers to consider the forms they used more carefully. In Lira's data, such was not always the case since language was recorded as it was produced online. Thus, the very nature of the two datasets separates the two studies, i.e., the two studies do not complement each other, but they deal with very different issues that are generated by the nature of their data.

[18] This is an important argument for this work and will be discussed in detail in Chapter 04. Its importance lies in the methodological application that the three subject persons must be analyzed separately from one another since it is agreed that each shows different patterning. The separate analyses will provide us with the opportunity to see how expression is realized within the same group of speakers, and what constraints are imposed in each person.

Indeed, by examining different datasets, the authors were investigating the same linguistic phenomenon under different sets of constraints. Paredes Silva, unlike Lira, demonstrated that the following factors condition the realization of expressed subjects in BP:

a) discourse continuity or connectedness;

b) emphasis;

c) ambiguity;

d) clause type;

e) distance;

f) position in the clause

g) person (referent);

h) animacy.

When compared to the findings presented in Lira, Paredes Silva innovates by elaborating a more detailed model of subject continuity. She proposed a system to codify referents based on their continuity in discourse, ranging from subjects that are very continuous[19] to subjects that are mentioned only (1993, p. 43). Evidently, she found that referents higher up in her discourse connectedness scale would show a disfavoring for overt subjects, whereas referents that fared lower in her scale tend to be expressed subjects, along the same lines as Lira, who only measure the continuity from the previous clause.

It is no surprise that Paredes Silva's results show such a strong tendency across her discourse connectedness continuum. Other scholars examining various dialects of Spanish have found a similar tendency for unexpressed subject to emerge in contexts of maximum discourse connectedness, while expressed subjects become more salient when the discourse

[19] Li and Thompson point that this kind of referent forms a "topic chain" (Li & Thompson, 1976).

connectedness is disturbed (Ávila-Shah, 2000; Bentivoglio, 1987; Cameron, 1995; Morales, 1997; Silva-Corvalán, 2001; Travis, 2005). This convergence of findings toward a similar conclusion encourages the observation that subject expression is not only a product of the inflectional system of a language[20], rather subject expression emerges in discourse when a certain number of constraints are met, one of them being discourse connectedness.

Silva (1996) examines the realization of 3sg pronominal/unexpressed mentions in variation with full NPs in the way they are realized in informal letters. She shows that old information is realized pronominally 37% of the time, is unexpressed 50% and is realized as a full NP 13%. Interestingly, though, thirteen percent of the data was realized as NPs that had already been introduced. These findings have important implications for the hypotheses developed here. Firstly, if 3sg unexpressed mentions are a result of the persistence (cf. Givón, 1983b) of the NP in discourse, then why are the NPs repeatedly used? Secondly, Silva's figures seem to suggest that a second or third mention of the referent would be realized as a pronominal form while further mentions would be unexpressed. Givón (1995, p. 79) suggests that a distance of more than three clauses is an adequate measure to determine whether or not forms are still part of the topic of the conversation, in Chafe's terms, whether or not a form is still in the front of the speaker's and the hearer's consciousness (1994, p. 30). Unfortunately, these measures are not provided by Silva, but their lack thereof does not overshadow the importance of her findings. This question of the persistence of referents and

[20] This conclusion has been noted in the generative literature on the basis of languages that show a rich inflectional system, e.g., German, that do not present a fully functional system of *pro-drop* (Huang, 1984; M. Modesto, 2000a). Moreover, the very nature of a language like Chinese, whose agreement system is very poor to license null subjects, still prolifically licenses null subjects.

their realization as either expressed or unexpressed subjects will be investigated in more detail in the dissertation.

In short, functional and variationist accounts of subject expression add to the literature the notion that form and meaning are not discreet, that is, when speakers make a choice between an expressed subject in place of a unexpressed subject there seems to be discourse and pragmatic reasons behind their choice. Such motivation is a product of the speaker's cognitive schemas that have been experienced over one's lifetime. In other words, everyday speech, or the speaker's experience with language shapes their representation of abstract structures, which in turn are divulged in their discourse. The studies in BP reviewed here showed the attempt of the researchers to capture and interpret the patterns that emerge from the speakers' discourses in an attempt to understand the phenomenon of subject expression. These results show similar patterning of pronominal and unexpressed subjects in their different data. However, not only their work, but more formal accounts focus on the variability at a more general level by examining all persons together, which is one of the major weaknesses of these studies. Thus, this study will contribute to the existing literature by offering a comparison of the three persons separately to observe what conditioning factors apply to each person.

3 METHODOLOGY

In this chapter I will describe the methodology used to collect the data and extract the tokens to be analyzed in this study, including defining the envelope of variation. I will also present a detailed description of each of the factor groups tested in the Variable Rule Analyses (VRAs), and the hypotheses that support the choice for each of them.

3.1 Overview of Variationist Methodology

In this theoretical framework, what matters is the relative frequency of co-ocurring linguistic forms (Cedergren & Sankoff, 1974; Sankoff, 1988b; Sankoff & Labov, 1979; Wolfram, 1993), or the frequency with which a certain structure occurs in discourse. In short, it is up to the researcher to (a) identify the linguistic phenomenon to be examined, in the case of this work, subject expression; (b) list the variables in competition which will serve as the dependent variable, here pronominal vs. unexpressed subjects; (c) raise hypotheses that encompass the systematic tendencies of the dependent variable; (d) operationalize the hypotheses through independent variables, or factor groups of both linguistic and/or social nature; (e) identify, collect and code the relevant data; (f) and finally, submit the coded data to statistical analysis and interpret the results obtained. I will discuss each of these steps further in this section and in the subsections that follow.

One of the central methodological questions within variationist theory consists of designing and defining mathematical models that are capable of associating adequately the relative weights, probabilities, with the several factors of each independent variable or factor groups to measure the influence that each factor exerts on the realization of one or another variable from the dependent variable. This is of importance because the factors of the several

independent variables occur concomitantly in the contexts or environments where the dependent variables are realized. More precisely, the possible number of contexts is a combination of the several factors of each independent variable. And, according to Labov, in order for one to formulate a set of rules, it is necessary to develop a methodology to quantify the factors, in a relatively small number, each showing a fixed weight, independent of the context where they occur (cf. Cedergren & Sankoff, 1974; Labov, 1972b).

Probabilistic models that calculate the relative effect of the independent factor groups based on observed frequencies were introduced in variationist research by Cedergren & Sankoff in 1974. Later, Rousseau & Sankoff present a new model defined as mixed or logistic, which is considered more appropriate to analyze variable phenomena. Discussions about the models used before this can be found in Rousseau & Sankoff (1978). This model has been used successfully and its description can be found in detail in the literature on variationist methodology (Cedergren & Sankoff, 1974; G. Guy, 1988; Naro, 1981, 1992; Sankoff, 1988b).

This statistical model, thus, posits that, in binary phenomena such as pronominal expression where you have two variables, pronominal or null, probabilities closer to a value of 1 favor the application of the rule relatively more than those closer to 0. In the case of the variable being examined in this study, the application of the rule is the realization of pronominal subjects, thus, the closer is the weight to 1, the greater the favoring of pronominal subjects and the disfavoring of unexpressed subjects. For example, when we observe the effect of the independent variable person on pronominal expression, i.e., application of the rule, we obtain the set of probabilities illustrated in Table 3. To interpret the results of the probabilities displayed in this table, each factor must be considered relative

to the others within the factor group. Thus, of 1sg, 2sg and 3sg subjects, that which most favors expression is 1sg (with a probability of .60), and that which least favors expressed subjects is 3sg (with a probability of .38). 2sg (with a weight of .50) favors expressed subjects less than 1sg, but more than 3sg.

Table 3. Hierarchy of constraints for PERSON.

Total *N*	8066				
% *expressed*	56.2				
Corrected *Mean*	.570				
		Probability	**% expressed**	**N**	**% data**
Person					
1sg		.60	64.7	3447	42.7
2sg		.50	53.4	1689	20.9
3sg		.38	47.7	2930	36.3
Range	22				

Total Chi-square = 1903.8191; Chi-square/cell = 1.6342; Log likelihood = -5105.385

This model, thus, is more adequate for examining linguistic variation because it operates with relative weights, or probabilities, rather than with simple percentages[21], and it quantifies the relative influence of each factor with regard to the dependent variable giving these factors their relative weights within a certain a factor group. What is more, this model is embedded with the notion that all factor groups are uniform, or orthogonal as Guy proposes (1993). This principle states that each factor group must be independent of the others in order for the model to work. However, linguistic reality sometimes defies this premise, and factor groups are indeed overlapping mechanisms in an analysis. Thus, there is the need to regroup or reanalyze overlapping variables into one in order for the model to conform to the principle of orthogonality (cf. Kay, 1978; Tagliamonte, 2006, p. 181).

[21] The percentages demonstrate the same relative rankings. However, the percentages do not take into account the interaction with the other factors, so the value of this approach is that it allows the analyst to identify the set of factors that jointly account for the variation in a statistically significant way.

In order to implement this mathematical model, David Sankoff developed a program called Varbrul, whose latest version, GoldVarb X (Sankoff, Tagliamonte, & Smith, 2005), is being used in this study. The program was written in FORTRAN and the way it functions is explained in detail in the following paragraphs.

GoldVarb not only calculates the relative weights of each factor within a factor group, or independent variable, but it also presents a statistical selection of factor groups that contribute to a significant analysis of the dependent variable. In short, the program provides the researcher with a hierarchical list of all the factor groups that significantly contribute to the realization of the application rule. This selection takes place at the significance threshold established at .05, which means that factor groups, or independent variables, which are chosen as significant are done so under a possibility of error of 5%. In other words, there is a 5% chance that the significant result has been obtained merely on the basis of statistical fluctuation, or error, and such result does not reflect a significant difference in the data.

A second element that influences the choice of a variable as significant is the *log likelihood*, which measures the degree of adequacy between the relative weights, or probabilities, and the observed frequencies, i.e. it measures the adequacy of the entire logistic model to the data in hand (Rousseau & Sankoff, 1978, pp. 60-61; Tagliamonte, 2006, p. 225). If the significance level of groups of values were to be the same, the program would then choose the group of values that has a *log likelihood* closest to zero. If there are still two groups with similar values and significance, then the program chooses the ones with the smallest number of factors as the significant one. A significance level of .000 is ideal because it indicates that the model fit perfectly to the observed data.

An important aspect of GoldVarb lies in the fact that it works with several levels of analysis calculating comparisons between the probabilistic values attributed to each of the variables entered in the analysis. At the first level, called zero, the program calculates the overall corrected average of the application rule when the effect of all the factors is neutral. This probability is called the *input*.

At the next level, the program calculates the probabilities of each of the variables in isolation in comparison to the *input*, then it attributes each variable a *log likelihood* and a level of significance and it finally chooses one of the variables to proceed to the next level.

Once the first variable is selected, the program executes another level of analysis whereby the selected variable is compared to each one of the other variables, separately, in pairs. Each pair is given a *log likelihood* and a significance level based on their probabilities until the program chooses a second variable that is more relevant from a statistical perspective.

Following this third level of analysis, the program compares the two variables that have been selected with the remaining, now in groups of three with the purpose of selecting another, a third, significant variable and it follows this procedure until it has selected all the significant variables. Thus, the number of levels present within an analysis is a function of the number of variables entered in the analysis.

This process described above, from level zero to level *N*, is called *stepup* and it is also carried out inversely, i.e. from level *N* to level 1, also called *stepdown*, to verify that all the variables not selected as significant in the *stepup* process are also eliminated in the *stepdown*.

These diverse levels of analyses are important because they allow the researcher the opportunity to verify with precision the interference between the variables, which can be the

result of overlapping coding or a natural interaction between the variables. When no interactions are observed, the probabilities are similar from level 1 to the last level of analysis. This is an ideal linguistic and mathematical circumstance, but also one that would undermine the necessity for such mathematical sophistication to account for variation in language.

In case of variables overlaying one another, the program attributes relative weights according to the statistical importance of each, based on, for instance, the balanced distribution of the data.

Thus far I have given an overview of the variationist framework that is going to be employed to understand the motivation between expressed and unexpressed pronouns in these data. As was mentioned earlier, variationist theory also provides tools to investigate change and variation. Here I will use the software package Goldvarb X (Sankoff, et al., 2005) which is capable of measuring the effect of several factor groups simultaneously to identify a set of factors that condition the linguistic variable, in this case subject expression.

3.2 Procedures

3.2.1 Corpus and Data

The data used for this study come from the corpus of oral educated Brazilian Portuguese (PORCUFORT) recorded in Fortaleza between the years of 1991 and 1994 (Monteiro, 1994a). The participants are all native speakers of Brazilian Portuguese born in Fortaleza whose parents were also raised in the city. This provides us with a relatively regionally homogeneous group of speakers.

The corpus was collected by Monteiro and a group of graduate students in the city of Fortaleza. The corpus was then transcribed and published (Monteiro, 1994a) and made freely

available for researchers of Brazilian Portuguese. After contacting Prof. Lemos Monteiro, I acquired all the audio files as well as their original transcriptions and digitized them in 2008. During the digitizing period I proofed one third of the corpus for accuracy of transcription, and found a high level of accuracy such that it was deemed unnecessary to check the remainder.

This corpus was chosen for two reasons. Firstly, it is highly homogeneous in terms of the level of education of speakers and region. Since every participant has a minimum of college degree, it is expected that higher rates of non-expression will emerge since traditional analyses claim it to be the preferred pattern among more educated speakers (C. Cunha & Cintra, 1985; Mary Aizawa Kato, 1999). Secondly, the corpus is divided in three distinct registers[22], namely two-party conversations, sociolinguistic interviews, and formal lectures. The sum of the three subsets amounts to approximately 500,000 words, of which approximately 25% was used in this study (to 117,685 words).

In order to circumvent any possibility of bias toward the data used, I randomly selected a number of transcripts as to represent (a) the three data subsets equally, (b) both genders, and (c) three distinct age groups[23], namely group I (22-35), group II (36-50), and

[22] There is some discussion in the literature in terms of the terminology to be used here (i.e. *genre* or *register*). I follow Silva-Corvalán's definition of register as linguistic varieties distinguished by their mode of communication (2001, p. 151) (see also Biber, 1995 for discussion).

[23] The grouping was adapted from Monteiro (1994b) respecting his stratification of the data according to these three groups. As will be seen later in the study, only the older group is in fact behaving differently in terms of the distribution of expressed subjects. The younger and middle-aged groups are behaving very similarly, whereas the older group shows a slight favoring for the unexpressed subjects. This difference, however, is not significant in any level of analysis for this data.

group III (51+) (Monteiro, 1994b). Then from each transcript the first one thousand words were discarded for they may represent more formal speech (Labov, 2001). After that, all occurrences of main verbs with their respective subjects were extracted and coded in Microsoft Excel for a number of factors that have been shown to have an effect on subject expression. The corpus distribution is documented in Table 4.

Table 4. Corpus makeup.

	Men			**Women**			**# of words**
	group I (22-35)	group II (36-50)	group III (51+)	group I (18-35)	group II (36-50)	group III (51+)	
Conversations	4	3	3	7	3	2	34,078
Interviews	3	2	1	2	2	1	48,429
Lectures	3	3	4	2	3	2	35,178
Total	26 speakers			24 speakers			117,685

The data obtained by these protocols consists of 18,810 tokens of finite declarative clauses with 1sg, 2sg, and 3sg subjects, and they were coded on their subject realization, that is whether the subject was realized or not. Each of these tokens were scrutinized to see whether they fit within the envelope of variation (see section 3.2.2), and after the appropriate exclusions were made, the remaining tokens were coded for the following eight independent linguistic variables, which are potential predictors of the probability of appearance of an over pronoun. The constraints for each of the variables are listed right below them.

a) PERSON
   - 1SG
   - 2SG
   - 3SG

b) VERB CLASS
   - COGNITION
   - POSSESSION
   - RELATIONAL
   - SPEECH

- OTHER

c) MORPHOLOGICAL IRREGULARITY
   - REGULAR
   - IRREGULAR

d) TAM
   - PRESENT
   - IMPERFECT
   - PRETERIT
   - FUTURE

e) CLAUSE
   - MAIN
   - SUBORDINATE

f) MODAL
   - PRESENT
   - ABSENT

g) POLARITY
   - POSITIVE
   - NEGATIVE

h) DISCOURSE CONTINUITY
   - SAME SUBJECT AND SAME TAM
   - SAME SUBJECT AND DIFFERENT TAM
   - DIFFERENT SUBJECT AND SAME TAM
   - DIFFERENT SUBJECT AND DIFFERENT TAM

These independent variables were selected, for the most part, because previous studies had found them relevant to the occurrence of overt pronouns (Cameron, 1992, 1996; Cameron & Flores-Ferrán, 2003; Duarte, 1993, 2000, 2003; Enríquez, 1986; Lira, 1982; Paredes Silva, 1993, 2003; Silva-Corvalán, 1982; Silveira, 2008; Travis, 2005, 2007).

3.2.2 *Defining the variable context*

In all human languages, spoken and signed, we can find examples of cases in which speakers have multiple ways of saying the same thing. Some variation is accidental and transitory; it may arise from the mechanical limitations of the speech organs, for instance, and may not be fully under the speaker's control. Other, more systematic variations represent options speakers may consciously or unconsciously choose (Coulmas, 2005). A choice between two

or more distinct but linguistically equivalent variants represents the existence of a linguistic variable, or a sociolinguistic variable.

Labov observes that to define the sociolinguistic variable, the researcher must first define the exact number of variables as well as establish the linguistic contexts where these variables occur (1972b, p. 121). By obtaining such variables and their contexts one can quantify each variable within a context and submit these values to a rule application. Naro asserts that the acceptance of such variable rules is just as valid as to accept rules that "force speakers to produce certain forms categorically" (1992, p. 17)[24]. In short, it can be argued that in the same way that there are categorical structures which, if violated, can generate agrammatical structures (e.g., in BP one cannot postpose the article to the noun), there are also conditions or variable rules that "work to favor or disfavor, variably and with specific weight, the use of one or of the other variable in each context" (Naro, 1992, p. 17)[25].

From a linguistic perspective, on one hand, the choice of a variable depends on a number of factor such as "features in the phonological environment, the syntactic context, discursive function of the utterance, topic, style […]" (Sankoff, 1988b, p. 984). From an extralinguistic perspective, on the other hand, factors such as sex, age, and social class may also condition the choice of variables by the speaker. Besides these different factors, the "interactional situation," Sankoff states, must also be taken into account in the study of variation (1988b, p. 984). Thus, one speaker can demonstrate evidence of variation in their speech by using one or the other form, or sometimes both in their speech as can be seen in

[24] "(…) que obrigam ao falante a usar categoricamente certas formas." My translation.

[25] "(…) que funcionam para favorecer ou desfavorecer, variavelmente e com pesos específicos, o uso de uma ou de outra das formas variáveis em cada contexto." My translation.

example (19) below where the speaker uses a pronominal 1sg subject in the first mention, followed by an unexpressed 1sg subject.

(19) ***Eu tenho*** *um compêndido de literatura brasileira do Coelho Neto do começo do século... num me* ***lembro*** *o ano.*
'**I have** the anthology of Brazilian literature by Coelho Neto from the beginning of the century... **(I) don't remember** the year'

(L3:281)

This discussion is of vital importance because the core of variationist work relies on the delimitation of the variable context, or the envelope of variation: the linguistic environments in which all the variants under consideration may occur. However, this definition does not exist without a debate. To circumscribe the envelope of variation, the researcher must carefully look not only at the contexts themselves and their behavior, but at the values these contexts represent in relation to the variation in study. Tagliamonte (2006), following Guy (1988), posits that contexts that occur at extremes (e.g., at 95% or at 5%) should not be included in any variable rule analysis for these contexts do not behave in the same way as the rest of the data in relation to the variable. They can be treated as the categorical in nature and should not be analyzed. Otheguy et al. (2007) go farther than Tagliamonte in that they establish the inclusion or exclusion of certain contexts based on their low or high variability rather than "one between absolutely variable and absolutely invariable environments" (Otheguy, et al., 2007, p. 776). Thus, the grounds for what can be included or not within the envelope of variation must be established by the degree of variability of the context.

In this study, I focus on declarative statements, and was thus faced with the task of identifying all cases where there exists variation between pronominal and unexpressed subjects.

To determine what falls within the envelope of variation, the data was first scrutinized for contexts where the alternation between pronominal and unexpressed subjects showed the least variability or no variability of any kind. Following the methodology laid out in Tagliamonte (2006) and Otheguy et al. (2007), these contexts were identified and are listed in Table 5 and they will be discussed in the subjections below.

Table 5. Data excluded from the analysis.

	N	Total
Exclusions		
Gerunds	277	
Infinitives	886	
Non-referential	486	
Post-posed	160	
Que as head	1174	
Repairs	134	4355
Truncated	76	
se as subject	92	
Researcher's speech	529	
Mismatched agreement	63	
Answers to questions	478	
Constructions		
Existential constructions	640	
é	311	
era	832	
comé	226	2887
é que	189	
quer dizer	165	
entende?	125	

viu?	219
seja	98
será	34
Other	48
Total	**6010**

3.2.2.1 *Gerunds and infinitives*

Infinitives and gerunds are not inflected for person, number, or tense[26], and just 5% (N = 58) of the time occur with an overt subject in these data. Therefore, they were not included in the analyses.

3.2.2.2 *Non-referential subjects*

The first group of exclusions is that of non-referential subjects, including generic referents and impersonal referents. These subjects are categorically unexpressed. Example (20) demonstrates this usage. These subjects are called non-referential because they cannot be inferred from the context, or are what traditional grammar in BP calls an indeterminate subject (Negrão & Müller 1996; Negrão & Viotti 2000), or they cannot be retrieved from the previous environment.

(20) Inf. - *aí Ø é*
já Ø é
na própria escola Ø é
onde cê tá trabalhando a reciclagem
'Then (it) is
just (it) is
at the school is
where you are working the recycling'

(L5: 284-287)

This lack of referential retrievability alongside its categorical occurrences with unexpressed pronouns places these clauses outside the envelope of variation.

[26] Tokens of the personal infinitive were very few (N = 26) and were collaped with present tense and were, therefore, included in the study.

Furthermore, there are two other types of constructions that can be analyzed as non-referential as well, because of their lack of a subject[27]. These two types are clauses whose verb denotes climatic activity, i.e. climate verbs, and verbs that denote the existence of something, i.e. existential verbs. These two types are briefly described below.

3.2.2.2.1 Climate verbs

A group of clauses excluded in this study corresponds to those in which the main verb refers to climate or time. Thus, all nature verbs such as *trovejar* 'to thunder', *chover* 'to rain', and *amanhecer* 'to dawn', just to name a few, were excluded.

3.2.2.2.2 Existential verbs

Three particular existential constructions, namely *faz* 'do-3sg' as in (21), *há* 'there be-3sg' as in (22), and *tem* 'have-3ps' as in (23). These three constructions alone amount to nearly 93% of all tokens of existentials[28].

(21) *Como profissional faz pouco tempo. Tenho apenas três anos como profissional.*
'As a Professional it has been some time. (I) have just three years as a professional.'
(I21: 296)

(22) *Então há até um poema de Camilo Peçanha.*
'So there is a poem by Camilo Peçanha.'
(L3: 186)

(23) *Tem cadeiras suficientes nas salas.*
'There are enough chairs in the rooms.'
(C47: 474)

[27] All these clauses represented under this label of exclusion carry verbal inflection for person, but any semantic, syntactic, or pragmatic subject is completely nonexistent. Thus, these clauses cannot be analyzed in the same form as any others present in this study.

[28] The remaining tokens of existential verbs consists of the verbs *acontecer* 'to happen' and *existir* 'to exist' in different TAMs, thus not forming any crystalized structured as the ones described in this section.

These forms represent constructions crystallized to perform the functions they do. Note that they do not agree with their complement. Thus, the formulaic and non-variable nature of these constructions means that they fall outside the envelope of variation.

3.2.2.3 *Post-posed subjects*

Another context that was excluded from the analysis was that of post-posed subjects, for two reasons. Firstly, by definition, post-posed subjects are categorically expressed. Secondly, the variation to be observed with these subjects occur categorically with 3sg whereby subjects can be realized pronominally, lexically, or unexpressed. As Lira (1982) points out, postposed subjects in BP are most likely to be subjects that denote new information, and since pronominal subjects are mostly representing old information, their occurrence in postposed position is very rare. Thus, examples (24) through (26) illustrate post-posed subjects with lexical subjects. In these examples the subject has been bolded and the verb underlined for ease of identification.

(24) *Aí depois <u>chegou</u>* ***umas roupas comprida***
'Then after some large clothes arrived'
(I12: 180)

(25) *Porque já <u>tinha saído</u>* ***a maioria do pessoal***
'Because most people had already left'
(I13: 458)

(26) *Aí <u>começa</u>* ***o período de treino***
'Then the training period begins'
(C45: 350)

Furthermore, the nature of post-posed subjects presents the pursuit of a different question from the one at hand. It is agreed upon that in BP post-posed subjects are used to introduce new referents in discourse (Maia 1998; Zilles 2000; Fernandes 2004). While this seems to be an appropriate function of such a change from the basic word order of the language, this is not the only use for post-posed subjects, especially pronominal ones, which

were minimally realized as expressed pronouns (8%, N = 13). It has been argued that pronominal post-posed subjects are not introducing new referents in discourse; rather, they are reintroducing older referents that have been dormant for a while (Lira 1982). Traditional accounts have postulated that pronominal post-posed subjects are a context of emphasis supporting their claim for this constraint in the expression of pronouns (Quicoli 1976; Barbosa et al. 2005). However, the number of tokens of such pronominal subjects was so rare in these data that they had to be excluded for they could not really be analyzed in any meaningful way.

3.2.2.4 *'Que' as head of a relative clause*

Relative clauses in which the head is the subject of the verb in the subordinate clause were not included in this study because they rarely occur with a resumptive pronoun as in (27).

(27) *O Ricardo que do primeiro ano, morava na Bahia.*
'Ricardo who during the first year lived in Bahia.'
(C116:787)

3.2.2.5 *Truncated utterances*

An utterance was considered truncated when the speakers either did not produce the verb, or did not complete the verb form as in example (28). Such tokens were excluded because it was not possible to identify all the necessary contextual factors (such as tense).

(28) *Aquele primeiro-ministro alguns anos atrás que ele se <u>enfor</u>--*
'That Prime Minister some years ago, he enfor—'
(L19: 358)

3.2.2.6 *Speech produced by one of the researchers*

Speech produced by the researcher was not considered in this analysis such as

(29) *Vamo falar aqui um pouquinho sobre o real qual a posição de vocês aí diante desse quadro econômico?*
'Let's talk for a little bit about your position in relation to this economic situation?'

(I28:347)

3.2.2.7 *Quotes from written material*

Quotes from written material were also excluded from the analysis. The principle at work here is very straightforward, since I am investigating the distribution of expressed and unexpressed subjects in oral discourse, it is not methodologically appropriate to incorporate quotations from written material into the analysis.

3.2.2.8 *Fixed constructions*

Constructions that occurred categorically with one or the other form were also excluded from this study. Examples are presented below and the constructions have been bolded for ease of identification. These two constructions occurred categorically with unexpressed subjects, thus their being excluded from the statistical analysis. However, their effect on the overall pattern is recognized and discussed along with others that were found to be frequent are discussed in section 7.5.

(30) *Ele tinha como como referência uma árvore lá... tá certo?* ***quer dizer*** *pisando... em solo...* ***quer dizer*** *a área totalmente seca...né?*
'He had a tree as a reference, right? **I mean** stepping on the soil, **I mean** the area is totally dried out, right?'

(L52: 462)

(31) *Porque você num conhecia você num ouvia falar e tal... então* ***é*** *aquele negócio.*
'Because you didn't know you didn't hear of it a so on… so (it) is that same old thing.'

(I52: 170)

3.2.3 *Operationalizing hypotheses as factors*

After excluding 6,010 tokens that do not fit in the envelope of variation, 2,252 tokens of non-human referents, and 2,482 tokens of full lexical occurrences, a remaining 8,066 tokens

coded for a series of factors, adapted from hypotheses and findings in the literature. These factors are morphological, syntactic, semantic, and pragmatic in nature, as well as social. I will now list all factor groups adopted in this analysis.

3.2.3.1 Person

The first morphological factor group to be considered in this analysis is that of the subject person. Since this study is only concerned with 1st, 2nd, and 3rd person singular, the three categories are straightforward.

Duarte (2003), among other scholars, has noted that the realization of pronominal subjects does not occur evenly across all persons of speech.

Furthermore, scholars in the literature on subject expression have continually showed that person is the strongest factor group to condition the patterning of expressed subjects (Barbosa, 1995; Barbosa, et al., 2005; Barrenechea & Alonso, 1977; Duarte, 1993, 2003; Mary Aizawa Kato, 1999, 2000; Lira, 1982; M. Modesto, 2000a, 2000b; Otheguy, et al., 2007; Silva-Corvalán, 1982; 2001 inter alia). The accepted hypothesis states that person functions as constraint on subject expression, with 1sg being that which most favors expression and the 3sg most disfavoring the expression.

For the purposes of this study, we are interested in the ranking of the singular forms in terms of their rates and probabilities of expression. Most studies concur that 1sg and 2sg subjects tend to probabilistically favor expression more than 3sg subjects. This is, thus, the working hypothesis of the study at hand. Based on the findings of Silveira (2007), 1sg and 2sg subject higher rates of expression than the rates of 3sg subjects. Therefore, it is the hypothesis here that 1sg and 2sg subjects will show higher rates of expression than 3sg subjects.

3.2.3.2 *TAM*

The Tense, Aspect, Mood of the main verb was also tested in this analysis. The different TAMs were coded according to the following coding list illustrated in Table 6. The column 'categories used in analysis' corresponds to the necessary collapsing of the TAMs based on their semantics and their patterning with the dependent variable. From now on, when referring to **TAM**, I will be referring to the categories presented in the first column.

Table 6. Tense-Aspect-Mood used in the analysis.

Category used in analysis	TAM	Example
Present	Present	*vamos, vá, falamos, falá*
	Present Progressive	*estamos indo, esta indoesta falando, estamos*
	Present Subjunctive	*vamos, vá, falemos, fale*
	Present Conditional	*iriamos, iria, falariamos, falaria*
Future	Analytic Future	*vai falar, vamos falar*
	Synthetic Future	*falaremos, falara*
	Imperfect Future	*ia ir, iamos ir; ia falar, iamos falar*
	Preterit Future	*fui ver, fui levar*
	Future Subjunctive	*falar, falarmos; for, formos*
Preterit	Preterit	*fomos, foi, falou, falamos*
	Preterit Progressive	*estive indo. estivemos indo; foi + GER*
Imperfect	Imperfect	*iamos, ia, falava, falavamos*
	Imperfect Progressive	*estava indo. estavamos indo*
	Imperfect Subjunctive	*fosse, fossemos, falasse, falassimos*
	Past Perfect	*havia ido*
Excluded	Infinitive	*ir/falar*
	Gerund	*falando*

It is worth noting that there all the TAMs showed variability in expression. This factor groups tests two hypotheses, namely the classic notion of ambiguity, which is observed in the different persons across the different categories for **TAM** (e.g., 2sg and 3sg are ambiguous in all TAMs, while 1sg and the others show ambiguity only in the **IMPERFECT**). Secondly, I am interested in examining how expression is realized across the different ways of framing discourse events in time. It is postulated here that the discourse framing of events will override morphological ambiguity in conditioning pronominal expression.

3.2.3.3 *Morphological irregularity*

This variable tests the hypothesis of whether the regularity of the verb affects the realization of pronominal subjects. Because irregular verbs are more marked, they are less likely to occur with pronominal subjects (Barddal & Eythórsson, 2003; Barddal, et al., 2011; Hay, 2001). This factor group has not been tested in BP and it focuses on the verbs themselves and their forms, which goes along with the main premise of this study which is to show that individual lexical items play a strong role in affecting the way more global syntactic patterns manifest in language.

3.2.3.4 *Verb class*

The semantic factors observed here are associated with the main verb. The taxonomy in use comes from Silveira (2007, p. 235), which is an adaptation of Dixon's taxonomy (2005) to suit the Brazilian Portuguese data. Table 7 below documents the values used in this study along with some examples for each verb class.

The rationale behind this factor group comes from the finding that subjects and verbs have a very strong bond, i.e. certain verb types tend to pattern with certain subjects. To illustrate this point, 1sg subjects co-occur more often with speech and cognition predicates while 3sg subjects appear more frequently with relational verbs. These patterns have been found not only for BP (Silveira, 2007), but for Colombian Spanish (Travis, 2006) and spoken American English (Scheibman, 2001).

Table 7. Categories of verb class used in the analysis.

	Description[29]	Examples
Motion	the subject is a Mover	*chegar* 'to arrive', *ir* 'to go', *sair* 'to leave', *entrar* 'to enter'
Perception	two core roles: a Perceiver and an Impression	*escutar* 'to listen/hear', *ver* 'to see', *olhar* 'to look'
Cognitive	two core roles: a Cogitator and a Thought	*achar* 'to think', *saber* 'to know', *entender* 'to understand'
Speech	Speaker, Addressee, and Medium	*dizer* 'to say', *falar* 'to speak', *chamar* 'to call'
Relational	establishes a relationship between two states or activities	*ser* 'to be', *estar* 'to be'
Possession	two core roles: an Owner and a Possession	*ter* 'to have'
Affect	two core roles: an Agent and either a Target or a Manipulator or both	*atingir* 'to hit', *chocar* 'to crash', *corrigir* 'to correct'
Giving	three core roles: a Donor, a Gift, and a Recipient	*dar* 'to give'
Rest	two core roles: a Rester and maybe a Locus	*ficar* 'to stay', *permanecer* 'to rest'
Other	verbs that did not fit in any of the above categories	*morrer* 'to die', *fumar* 'to smoke', *vencer* 'to win', *operar* 'to use'

Given that subjects tend to co-occur with greater frequency with particular predicates than with others, it is hypothesized that expressed and unexpressed subjects also show different distribution among different predicates. This is in part based on the findings for Spanish and Portuguese that show that rates of expression vary according to semantic class (Bentivoglio, 1987; Enríquez, 1984, 1986; Monteiro, 1994b; Silva-Corvalán, 1982, 1994, 1997, 2001; Travis, 2005, 2007). Except for studies that addressed this factor group in examining one person at a time, this study is unique in that it tests this hypothesis in comparing the three persons and examining how these classes of verbs affect the rates of expression for each person.

3.2.3.5 Clause type

Bybee (2002a) argues that main clauses are more innovative, whereas subordinate clauses tend to be more conservative and retain older patterns. Thus, the hypothesis is that

[29] The descriptions provided here follow Dixon's descriptions for each verb type. These descriptions are, in turn, semantically based to capture the relationship between arguments as part of the core meaning of the predicate.

main clauses would be more advanced in the change and thus show a higher rate of expressed subjects than subordinate clauses. This is exactly what is found in Silveira (2008) for 1sg subjects in that main clauses showed a much higher rate of expressed subjects, whereas subordinate clauses showed an almost categorical favoring for unexpressed subjects.

Bybee's hypothesis is supported by the way subordinate clauses evolve in the course of language change. Deutscher (2000) and Heine and Kuteva (2007) show that what we conceive today as subordinate clauses were once main clauses. Thus, these authors conclude that some of the patterns that remain in languages are a result of existing at a moment of that language's life and being trapped in the subordinate clauses structures that remain, although this has not been shown for Portuguese. It is possible that it is so if we assume that this is a common pattern in language change in general. Moreover, the syntactic structure of clauses of the vulgar Latin variety that changed into Portuguese later on showed a preference for unexpressed subjects (Posner, 1996). Such pattern can still be seen in subordinate clauses in BP.

In this study main clauses encompass both main clauses and also coordinate clauses because in preliminary analyses they pattern in the same way. The same is true for subordinate clauses which is seen as a unified category. This category initially consisted of several different factors (e.g., relative, adverbial, etc.), however, these sub-categories did not yield any significant differences in the findings obtained for this factor group.

Interestingly, other researchers have found very different results from the ones expected here. In fact, their results contradict our hypothesis. It has been found that subordinate clauses, especially relative and adverbial ones tend to favor pronominal subjects (Duarte, 2003; Ferreira, 2000; Lira, 1982; Monteiro, 1994b). What we expect to show in this

study is that the change continues and these results no longer hold to explain the phenomenon.

3.2.3.6 *Intervening element*

This factor group tests the hypothesis of whether the presence of an intervening element affects pronominal expression in BP. The rationale behind this factor group lies in the premise that subjects and verbs that tend to form a unit will have their constituency fragmented by some kind of intervening material. Consider (32) below. As can be seen when compared with (33), the form *eu já sei* 'I already know' in (32) does not function as a discourse marker, as has been argued for the form in (33). Furthermore, it appears that the presence of the adverb favors the expression of the pronominal form *eu* 'I'. Thus, it is clear that intervening elements do play a role in disrupting the structure of the more formulaic construction. Such a finding demonstrates the importance of this factor group.

(32) *Eu **já** sei o curso que eu quero, é esse aqui*
'I **already** know the class I want, it's this one'

(C47: 211)

(33) *A:* *Depois eu tenho também dicionário da Bíblia... que até um... um amigo meu o pastor S. de Cuba que me deu... aquele... que eu entrevistei*
B: ***Sei.***
A: *Que eu fui fazer pesquisa.*
'A: Besides I also have the Bible dictionary ... which ... a friend of mine, pastor S. Cuba gave it to me... that one .. who I interviewed
B: **(I) know**
A: When I was researching'

(C33:732)

3.2.3.7 *Polarity*

Silveira (2006) found that 1sg subjects tend to be left unexpressed more often in negative statements, which is in opposite direction of effect as that found in Duarte (1993). This

disagreement in findings leads to the need for us to explore this variable further and examine its effect on the realization of subject expression in BP.

3.2.3.8 *Discourse continuity*

The factor groups involved with discourse continuity of the referent represented by the subject will allow us to test for the hypothesis that adjacent syntactic forms tend to be isomorphic both in the form of their subject as well as in their verbal TAM. Furthermore, traditional analyses have argued that pronominal expression in BP is an outcome of speakers' intentions to disambiguate the subject of the immediately preceding clause. Thus, the hypothesis at study here is that the subject of the immediately preceding clause may either favor an unexpressed mention, supporting the argument of traditional analysis.

As far as distance and persistence are concerned, the hypothesis, put forth by Givón (1983b; 2001), is that topics that persist longer in discourse tend to become more attenuated in their linguistic form, thus we would expect unexpressed subjects to represent more persistent topics. As a starting point, persistence will be measured in terms of distance in clauses from the first to the last mention of the same referent up to twenty clauses.

The model was based on the following coding criteria adapted from Givón (1983b) Paredes Silva (1993), however, demonstrated that it is not just subject continuity, but a broader notion of discourse continuity that affects subject expression … :

- **Same subject and same TAM**

(34) *Inf. 2 - /cê já deve ter ouvido falar naquele... <u>Robert Lado</u>...*
Inf. 1 - uhn
Inf. 2 - o Lado ele criou uma metodologia muito interessante...
é::
é::...
é uma metodoloGIa
em que o professor já pode... PREVER o erro
que o aluno vai dar dentro da língua estrangeira...
por exemplo... você pega ah:: em nosso português

você...ele <u>ele es/... enSIna</u> comparando a língua... estrangeira o inglês
com a língua materna do {aluno o português...
então <u>ele faz</u> sempre um estudo
'Inf. 2 – you must have heard of… Robert Lado…
Inf. 1 – uhum
Inf. 2 – Lado he created a very interesting methodology…
yeah…
yeah…
(it) is a methodology
in that the teacher can… predict the error
that the student will produce in a foreign language
for example… you take Portuguese
you… he teaches it comparing the language… foreign with English
with the mother tongue of the student Portuguese
then he always does a study'

(C47:652)

This factor codes for the subject and TAM of one clause as being the same as those of the previous clause. As can be seen here, the last two underlined clauses have the same subject and the same TAM. Non-referential subjects were not considered as intervening clauses, that is, only clauses that had a referential subject were considered when looking back.

- **Same subject and different TAM.**

(35) *Inf. 2 - ah Tânia pois é bom...*
ah:: ai meu Jesu::s... dessas aula que <u>assisti[1]</u>...
o... concurso num <u>gosto[2]</u> nem de falar nesse assunto...
<u>fico[3]</u> calada...
' Inf. 2 – ah Tânia so
my goodness ... of all these courses (I) <u>attended[1]</u>
the contest (I) don't even <u>want[2]</u> to talk about it…
I <u>remain[3]</u> quiet…'

(C116: 377)

In this example, we see that clause 2 was coded as having the same subject and a different TAM from clause 1. The former has a verb in the preterit and the latter has a verb in the present. Clauses 2 and 3 have the same subject and same TAM.

- **Different subject and same TAM.**

(36) *Inf. 2 - a Erinalda disse*
é Vera
/tá muito difícil
resolver os pro-blemas
porque... a Secretaria /tá saBENdo[1]
os professor que estão faltando[2]
'Inf. 2 – Erinalda said
yeah, Vera
(it) is really difficult
to solve the problems
because the secretary is aware[1]
the teachers who are missing classes[2]'

(C116:600)

In this example, clauses 1 and 2 have the same TAM but even though they adjacent to one another, they have different subjects.

- **Different subject and different TAM**

(37) *ah burguesia*
como você me perguntou[1]
ela influencia[2] realmente na:: na produção literária...
'ah burgeousie
as you asked[1] me
they really influence[2] literary production…'

(L36:272)

In this example, we see that clauses 1 and 2 have different subjects and different TAM. The former is in the preterit and the latter is in the present.

3.2.3.9 Summary

With the aim of discovering the set of factor groups which jointly account for the largest amount of variation in a statistically significant way (Sankoff, 1988b), all factor groups were considered individually and together in multivariate analysis using GoldVarb X (Sankoff, et al., 2005). I now turn to the discussion of these results.

4 RESULTS OF OVERALL VARIABLE RULE ANALYSIS

This chapter presents the results of the statistical analyses of the effect of linguistic factors on subject expression. Several Variable Rule Analyses (VRAs) were performed to assess the impact of a number of factors on the likelihood that subjects would be realized pronominally. The full model contained the dependent variable, **EXPRESSION**, and eight independent variables, namely **TAM**, **VERB CLASS**, **CLAUSE TYPE**, **MORPHOLOGICAL IRREGULARITY**, **MODAL**, **POLARITY**, **DISCOURSE CONTINUITY**, and **SUBJECT PERSON**. The full model containing all predictors was successfully able to predict the conditioning environments most likely for expressed pronouns to be realized. This model establishes seven of the eight independent variables as statistically significant in the conditioning of pronominal subjects, namely **VERB CLASS**, **CLAUSE TYPE**, **PERSON**, **DISCOURSE CONTINUITY**, **TAM**, **POLARITY**, and **MODAL**.

While it is important to have a general picture of how pronominal expression behaves linguistically in the language, it is also critical to observe the individual patterns of each person separately. The comparison between separate analyses will offer us the unique opportunity to examine the variation from a global to a more local perspective, which can elucidate our understanding of the conditioning factors governing each person within this more global pattern. So, after discussing the overall findings from these data, I will report on separate analyses performed for each person and how they differ from the general analysis presented.

This chapter will be divided as follows. Firstly, I will briefly discuss the general results for the analysis of all persons and predictors combined. In the second section of this chapter, I will discuss separate analyses for each person and how the factors affect the

distribution of expressed subjects for each person. Finally, I will end with a discussion of the overall patterns observed in the several analyses presented.

4.1 Factor Groups selected as statistically significant

The full model containing all predictors was statistically significant and the results are illustrated in Table 8 where the factor groups are organized by their effect, from strongest to weakest, in conditioning the expression of subjects. Before we begin to discuss the results presented in Table 8, let us explain the table and the way the results are presented. In Table 8 and the subsequent tables, the 'input' indicates the overall likelihood that the variant – a pronominal subject – will occur. In the first column, the numbers represent the probability (or factor weight) that each factor contributes to the occurrence of the variant: the closer to 0, the less likely that pronominal subjects will occur with that factor and the closer to 1, the more likely that it will be. The range provides an indication of the relative strength of each group of factors in the analysis. In these results, **VERB CLASS** has the strongest effect, with strong effects also for **CLAUSE TYPE** and **PERSON**, while **DISCOURSE CONTINUITY**, **TAM**, **POLARITY** and **MODAL**, though significant, are relatively weak. The second column shows the percentages of pronominal subjects, the third column the total number of tokens in each factor, and the fourth column the percentage of the data each factor makes up. I will be focusing on the factor weights of the first column, which indicate the constraint hierarchy, or direction of effect.

A total of 8,066 tokens were included in this analysis distributed across 471 verb types. Pronominal subjects account for 56% of the data (N = 4530) and unexpressed subjects account for the remaining 44% (N = 3536). These tokens were submitted to a multivariate analysis and the results are documented in Table 8.

Table 8. Multivariate analysis of the factors that contribute to a statistically significant effect on the realization of pronominal subjects.

		Probability	% expressed	N	% data
Total *N*	8066				
% *expressed*	56.2				
Corrected *Mean*	.570				
Verb class					
Possession		.64	68.7	847	10.5
Speech		.59	63.6	1102	13.7
Other		.55	60.1	3144	39.0
Relational		.51	53.3	754	9.3
Perception		.38	45.7	488	6.1
Cognition		.32	42.2	1731	21.5
Range	36				
Clause type					
Subordinate		.70	76.6	1174	14.6
Main		.46	52.7	6892	85.4
Range	24				
Person					
1sg		.60	64.7	3447	42.7
2sg		.50	53.4	1689	20.9
3sg		.38	47.7	2930	36.3
Range	22				
Discourse continuity					
Diff Subj		.54	59.3	4334	53.7
Same Subj & Diff TAM		.48	55.7	1674	20.8
Same Subj & Same TAM		.44	50.0	2058	25.5
Range	10				
TAM					
Imperfect		.56	64.0	999	13.2
Preterit		.51	59.4	1695	22.4
Present		.48	52.4	4862	64.3
Range	08				
Polarity					
Affirmative		.51	56.8	7136	88.5
Negative		.46	51.6	930	11.5
Range	05				
Modal					
Absent		.51	56.3	6688	82.9
Present		.47	55.5	1378	17.1
Range	04				
Morphological irregularity					
Regular		[.51]	56.3	3790	47.0
Irregular		[.49]	56.0	4276	53.0
Range	*n.s.*				

Total Chi-square = 1903.8191; Chi-square/cell = 1.6342; Log likelihood = -5105.385

Table 8 illustrates three levels of effect on these data, at one level it can be seen that **VERB CLASS** is the factor group that shows the strongest effect in conditioning variable subject expression, with a Range of 36, 1.5 times as high as the next strongest factor group; then we observe a second level of effect of **CLAUSE TYPE** and **PERSON**, which show about 2 times higher effect than the next group; and finally, a third level where **DISCOURSE CONTINUITY**, **TAM**, **POLARITY** and **MODAL** exert an effect, but a much weaker effect than in the previous two levels. Moreover, **VERB CLASS** being the strongest factor group in conditioning pronominal expression suggests that there is a strong lexical effect, whereby individual verbs or classes of verbs demonstrate preferences that override the overall syntactic pattern. In the following subsections I will discuss each of the significant factors in detail and offer explanations for why they have been chosen. However, as will be shown in Chapter 5, these results must be taken very cautiously because each person behaves differently in the ways these factor groups condition the variable. So, the main purpose of this chapter is twofold, (a) it is going to show that these independent variables hold up in their predicted effects in these data, and (b) that analyzing subject expression in such a generalized way may not offer the linguist with the opportunity to examine what is really happening at a more detailed level.

4.1.1 Verb class

Recall from section 3.2.3.4 that several researchers have argued for the notion that predicates and subjects tend to co-occur often enough that they appear to be bonded. In other words, certain predicates are probabilistically more likely to occur with specific subjects than they are with others. Departing from such premise, this factor group seeks to understand what these bonds are and how they affect the realization of pronominal subjects.

VERB CLASS is the strongest factor in predicting the occurrence of a pronominal subject in these data. While other studies have showed the effects of **VERB CLASS** on subject expression, this is the first study to actually have observed this factor group to have such a strong effect. For ease of readability, the hierarchy of constraints observed in Table 8 are reproduced here in Table 9.

Table 9. Hierarchy of constraints for verb class.

Total *N*	8066			
% *expressed*	56.2			
Corrected *Mean*	.570			
	Probability	**% expressed**	**N**	**% data**
Verb class				
Possession	.64	68.7	847	10.5
Speech	.59	63.6	1102	13.7
Other	.55	60.1	3144	39.0
Relational	.51	53.3	754	9.3
Perception	.38	45.7	488	6.1
Cognition	.32	42.2	1731	21.5
Range 36				

It can be observed that there is a clear division between the verb classes in regards to their effect on pronominal expression. **POSSESSION** predicates show the strongest favoring to co-occur with pronominal subjects. Then we see **SPEECH** and the category of **OTHER** favoring pronominal expression with **RELATIONAL** predicates tailing behind. It must be noted that **RELATIONAL** predicates are considered to favor expression in terms of their effect in relation to the other factors within this group, i.e., when compared to **PERCEPTION** and **COGNITION** predicates, we can see that **RELATIONAL** ones indeed favor pronominal subjects more. On the other hand, **PERCEPTION** and **COGNITION** predicates highly disfavor the realization of pronominal subjects, with the latter showing the highest probability against the realization of the dependent variable. This is striking because **COGNITION** verbs account for almost a

quarter of the data (21.5%), thus its effect as a class appears to be responsible for the retention of unexpressed subjects.

Finally, the patterning observed with **RELATIONAL** and **PERCEPTION** verbs, however, must be taken with care because they only account for a small portion of the data (9.3% and 6.1% respectively).

A number of studies have reported that subject expression interacts with **VERB CLASS** (Bentivoglio, 1987; Enríquez, 1984; Silva-Corvalán, 1994; Torres Cacoullos & Travis, 2010; Travis, 2005, 2007). These studies have found that **COGNITION**[30] predicates are strongly correlated with pronominal expression, especially with 1sg subjects. **RELATIONAL** verbs, on the other hand, are usually found to disfavor the realization of pronominal subjects in their data,

Firstly, the finding that **RELATIONAL** predicates favor the realization of pronominal subjects is in agreement with Enríquez (1984, p. 240) and it is also in accordance with Ashby and Bentivoglio (1993, p. 63) who noted that subjects of relational predicates behaved differently from subjects of other intransitive verbs, in that they tend not to occur as full Noun Phrases in both Spanish and French. On the other hand, this finding is in disagreement with Dutra (1987) who observes that these predicates favor the omission of subjects in her data[31]. Secondly, **SPEECH** predicates have also been found to favor the realization of pronominal subjects (Travis, 2005, 2007). Thirdly, **COGNITION** predicates have been widely associated with pronominal 1sg subjects in Spanish (cf. Silva-Corvalán, 1982, 1994, 2001;

[30] In other studies these predicates have been dubbed 'psychological', 'mental', etc. However, all these different labels will fall into the rubric of cognition that is being used here. The same adaptation will be used for the other classes of verbs.

[31] This may be due to dialectal differences between these data and Dutra's.

Travis, 2005, 2007), however, this is not the case in these data where these predicates highly disfavor the realization of pronominal subjects. Silva-Corvalán noted that verbs that express the opinion of the speaker favor explicit subjects more than other verb classes (Silva-Corvalán, 1994, p. 162). In the case of the first person, the high use of explicit subjects with these verbs has been attributed to the epistemic role such constructions play (Scheibman, 2001; Thompson, 2002). So, it is unexpected that these data are showing a different patterning than those observed in other studies. In sections 5.2.1, 5.3.1, and 5.5.2, I will offer explanations for why this unexpected patterning has emerged in these data. In these sections I will emphasize the role that verb *saber* 'to know', or to be more precise, the constructions *sei* '(I) know', *não sei* '(I) don't know', and *sabe* '(you) know' play in the behavior of this entire class and why it disfavors expression.

4.1.2 Clause type

Scholars have remarked that **SUBORDINATE** clauses, especially relative and adverbial clauses, tend to favor pronominal expression (Duarte, 1993; Ferreira, 2000; Lira, 1982; Monteiro, 1994b). These findings do not support the hypothesis that **SUBORDINATE** clauses are the locus of unexpressed subjects as proposed by findings in studies that follow a generative framework (Mary A. Kato, 1996). Studies of a functional nature have also pointed to this favoring of **SUBORDINATE** clauses for unexpressed subjects (Silveira, 2008). These conflicting findings suggest that **CLAUSE TYPE** may not be a truthful predictor of pronominal expression or that there may be something else at play that is causing this factor group to act randomly.

CLAUSE TYPE is the second strongest factor group to condition the realization of pronominal subjects in these data. The findings observed in Table 8 are reproduced here in Table 10.

Table 10. Hierarchy of constraints for clause type.

Total *N*	8066				
% *expressed*	56.2				
Corrected *Mean*	.570				
		Probability	**% expressed**	**N**	**% data**
Clause type					
Subordinate		.70	76.6	1174	14.6
Main		.46	52.7	6892	85.4
	Range 24				

In these data **SUBORDINATE** clauses favor pronominal subjects and **MAIN** clauses disfavor them. This finding is most surprising since it is argued that pronominal expression is a newer developed in the language and it would be expected to be favored by **MAIN** clauses, which according to Bybee and Thompson are more innovative in the syntactic structures they realize. What may be a possible explanation, and at this point it is just a conjecture, is that these data are an example of the language at a stage where pronominal expression and preferred pattern, or its innovative realization, has already evolved from the main clause to the subordinate clause as it is expected in language change.

4.1.3 Person

As was stated earlier in 3.2.3.1, the working hypothesis for this factor group is that the three persons would show different rates of pronominal expression. Most importantly, this difference in rates of expression suggests that these different persons are subject to different conditioning, different constraints, in short, differences which are lost if these persons are collapsed in an analysis.

This factor group is the third strongest following **VERB CLASS** and **CLAUSE TYPE**. It is interesting to note that this is not among the strongest factor groups here, though it has been found to be so in other studies (Duarte, 1993, 2003; Lira, 1982; Otheguy, et al., 2007; Silva-Corvalán, 1994, 2001). The findings for this factor group reported in Table 8 are reproduced here as Table 11 where we can see that both 1sg and 2sg subjects favor pronominal expression while 3sg subjects disfavor it. This finding replicates those observed in the literature on subject expression.

Table 11. Hierarchy of constraints for person.

Total *N*	8066			
% *expressed*	56.2			
Corrected *Mean*	.570			
	Probability	**% expressed**	**N**	**% data**
Person				
1sg	.60	64.7	3447	42.7
2sg	.50	53.4	1689	20.9
3sg	.38	47.7	2930	36.3
Range 22				

It must be noted that the distribution of each person in these data corresponds to findings observed by Scheibman (2001) for spoken American English in that 1sg subjects are more frequent than 3sg animate subjects, which in turn are more frequent than 2sg subjects in discourse. This is an important observation because if we are witnessing a process of change from variable to obligatory expression and 1sg subjects are the most frequently occurring subject person in discourse, then it leads us to believe that they are leading the change toward pronominal expression. This higher percentage of occurrences of *eu* 'I' has been explained as a consequence of the egocentric nature of verbal communication: by explicitly referring to himself, the speaker fulfills the pragmatic need to keep himself overtly present in the verbal interaction (Morales, 1986).

While the results for **PERSON** are fairly clear and in full agreement with those observed for BP and several varieties of Spanish, they are poignant for the argument we want to make in this work, namely that the three persons must be examined separately and comparably so as to allow us to interpret the patterns of change more efficaciously. If we observed the pattern demonstrated in Table 11 carefully, we can see that 1sg really favors pronominal expression, 2sg subjects slightly favor pronominal expression, and 3sg strongly disfavors it. Thus, these results are very suggestive that the three persons are behaving differently. This hypothesis will be explored in detail in Chapter 5 where the three persons will be subject to separate analyses using the same factor groups under discussion here.

4.1.4 Discourse Continuity

Recall that the model developed to test this factor group was based on the relationship between subjects and TAMs and their relationship with the previous clause. These relationships were arranged in four different levels of sameness and differences between subjects and TAMs. Levels one (Same subject and same TAM) and two (Same subject and different TAM) show a relationship of same referent and same and different TAMs. Levels three and four depict the relationship between different referents and same and different TAMs. In preliminary VRAs levels three (Different subject and same TAM) and four (Different subject and different TAM) patterned so similarly that they were collapsed in the remaining of the analysis as the factor 'different subject'.

This factor group has been shown to be a major determinant of pronominal subjects in Spanish and Portuguese (Ávila-Shah, 2000; Paredes Silva, 1993). The findings reported in Table 8 and reproduced here in Table 12 agree with previous research in that subjects and TAMs that are more continuous tend to be correlated with clauses without an expressed

pronoun, while less continuous referents and TAMs favor the realization of pronominal subjects.

Table 12. Hierarchy of constraints for discourse continuity.

Total *N*	8066			
% *expressed*	56.2			
Corrected *Mean*	.570			
	Probability	**% expressed**	**N**	**% data**
Discourse continuity				
Diff Subj	.54	59.3	4334	53.7
Same Subj & Diff TAM	.48	55.7	1674	20.8
Same Subj & Same TAM	.44	50.0	2058	25.5
Range 10				

This factor group tests several hypotheses concomitantly. Firstly, it tests whether or not switch reference plays a role in the expression of pronominal subjects in BP. The results observed here adhere to this claim and to previous findings (Cameron, 1992, 1994, 1995, 1996; Cameron & Flores-Ferrán, 2003). In contexts of switch reference there is a tendency for subjects to be realized pronominally, while the inverse is true.

Secondly, the notion of continuity was measured by coding for changes between previous referents and TAMs. Again, the results show that a change in TAM, but not in subject, is enough to raise the probability of expressed subjects occurring by .04 points, nearly 10%. These findings are most illuminating in light of what is known about the pragmatics of pronominal subjects. However, as I will demonstrate later in Chapter 5 that these findings do not hold true for all persons and they do not demonstrate the same patterning.

4.1.5 TAM

This factor group measured two distinct hypotheses for the conditioning of pronominal subjects in BP. The first hypothesis assesses the effect of ambiguity in conditioning these

subjects. as can be seen in Table 13, this hypothesis holds only for a comparison between the three subjects in one tense, the **IMPERFECT**. In this tense, all three singular forms have the same morphological inflection, thus being ambiguous in discourse. The favoring of this tense toward the realization of pronominal subjects suggests that ambiguity plays a role in pronominal expression.

While we acknowledge that there is an effect with the **IMPERFECT**, the notion of ambiguity does not behave in a consistent manner across the possible ambiguous TAMs. The other two tenses, the **PRETERIT** and the **PRESENT**, show ambiguity only across 2sg and 3sg subjects, and they do behave differently from the **IMPERFECT**. The former very slightly favors pronominal expression but the effect is weak compared to the **PRESENT**, which disfavors pronominal expression. Thus, while the argument for morphological ambiguity may have some reality in discourse, it is not an absolute across all possible ambiguity scenarios, which leads us to interpret the results for TAM in light of a more functional account.

Table 13. Hierarchy of constraints for TAM.

Total *N*	8066			
% *expressed*	56.2			
Corrected *Mean*	.570			
	Probability	**% expressed**	**N**	**% data**
TAM				
Imperfect	.56	64.0	999	13.2
Preterit	.51	59.4	1695	22.4
Present	.48	52.4	4862	64.3
Range 08				

Table 13 shows that both TAMs with past reference, namely the **IMPERFECT** and the **PRETERIT**, favor expression, with the former showing a slight stronger effect.

A functional account to explain the effects of TAM relies on Silva-Corvalán's proposal that the rates of expression correlate with the function of the TAMs, she claims that the present and the preterit are factual, assertive tenses that place events in the foreground, while the imperfect is a backgrounding tense that is less assertive and non-factual (Silva-Corvalán, 1997, 2001). Thus, instead of attributing the conditioning of the variable to ambiguity resolution, the function of these tenses in discourse become fundamental in clarifying the nature of this phenomenon.

As I will show later on in this work, this argument can explain the patterning that we observe in this overall analysis, but it does not hold up when examining each subject separately. What can really be observed is that TAM is intrinsically connected with the previous realization, here coded as **DISCOURSE CONTINUITY**. What we will see is that it is not just ambiguity or the function of the TAM in discourse, but the fact that TAMs that tend to be repeated across clauses tend to correlate more with pronominal subjects then other TAMs that do not repeat across clauses as often, and thus do not correlate with pronominal subjects as much.

4.1.6 Polarity

As with presence of **MODAL**, **POLARITY** also tests the hypothesis of a patterning between subjects, predicates, and possibly negation markers. Duarte reported that pronominal subjects are unlikely to occur in negative statements in main clauses (Duarte, 1993). Thus, including this factor group in the analysis will not only allow us to test the notion of the connection between subjects and predicates but also how the syntactic organization of a clause contributes to the conditioning of pronominal subjects.

While **POLARITY** has been selected as a significant factor group in the conditioning of pronominal subjects, it is a weak predictor (range = 05) as compared to others such as **DISCURSE CONTINUITY** and **CLAUSE TYPE**. Because of this weak effect, I consider **POLARITY** to be a marginal predictor in that it is likely to be overridden by the others if their contexts present themselves. To put it in another way, **POLARITY** is likely to affect pronominal expression if and only if **VERB CLASS**, **CLAUSE TYPE**, and **DISCOURSE CONTINUITY** fail to do so. The results presented in Table 8 are reproduced in Table 14 below. It can be seen that affirmative statements favor pronominal subjects while negative ones disfavor it, supporting the findings reported in Duarte (1993).

The finding that affirmative clauses favor the occurrence of pronominal subjects suggests that the basic sentence type in BP, namely a declarative, affirmative, **MAIN** clause is also one with an expressed subject when such is a pronoun. Since these sentence types are deemed more frequent than other types (Lambrecht, 1994, 2001), they may also be contributing to the spread of the pattern of pronominal expression throughout the language.

Table 14. Hierarchy of constraints for the factor group polarity.

Total *N*	8066			
% *expressed*	56.2			
Corrected *Mean*	.570			
	Probability	**% expressed**	**N**	**% data**
Polarity				
Affirmative	.51	56.8	7136	88.5
Negative	.46	51.6	930	11.5
Range 05				

Negative statements, as opposed to affirmative ones, are less likely to occur with a pronominal subject. This is possibly due to the fact that negative statements tend to convey given information (Fillmore, 1975; Givón, 1976; Givón, 1984, 1987; Givón, 2001; Givón,

1983c). Example (38) illustrates this point in that the subject 'ele' is presupposed in the negative statement. Fillmore argues that in order for a negation to be made, the content of the assertion needs to be presupposed by both the speaker and the interlocutor (1975).

(38) *Ele percebeu que ela tinha ido por causa dele, olha, começou a chorar e não conseguiu mais fazer a prova de matemática.*
'he realized she was there because of him, look, he started to cry and couldn't do the math test anymore.'
(C07: 566)

Another possible explanation, which will be entertained later in this work, is that there is a strong patterning of subjects, negative markers, and predicates where the subjects are not present in the clause. These constructions, I will suggest, contribute to the patterning observed in the table above.

4.1.7 Presence of a Modal

The hypothesis tested by this factor group follows from the same premise that verb class tests, that is, linguistic forms are bound to one another by their rates of co-occurrence. Thus, it is expected that modals show a pattern of co-occurrence with subjects, as do main verbs. However, it must be noted that this pattern of co-occurrence may not be as straightforward as it is with predicates and subjects. In the case of modals, there may not only be a bond of co-occurrence between the subject and the modal, but there may also be a bond between the modal and the predicate. Thus, we must account for a three-way bond connection between subject, modal and predicate.

The pattern observed here follows that a clause which does not have a modal verb tends to be realized with pronominal subjects, while the inverse is also true for clauses with modal verbs. Even though these findings suggest a possible correlation between the presence of a **MODAL** and pronominal expression, they follow the same logic presented for **POLARITY**,

i.e., that these results are suggestive at most given the weak effect observed through its range (04) in comparison to other factor groups such as verb class.

Table 15. Hierarchy of constraints for presence of modal.

Total *N*	8066			
% *expressed*	56.2			
Corrected *Mean*	.570			
	Probability	**% expressed**	**N**	**% data**
Modal				
Absent	.51	56.3	6688	82.9
Present	.47	55.5	1378	17.1
Range 04				

4.2 Discussion

The results observed in this analysis of the data for the three persons combined follow very similar patterns observed in previous studies of BP and several varieties of Spanish. It is recognized here that this is in part a result of the initial selection of factor groups to be used in the study. Since they are the same factor groups that have been found to have an effect on subject expression, these results are predictable.

As was mentioned earlier, the results can be grouped in three different levels based on their magnitude of effect. At level one we have **VERB CLASS**, **CLAUSE TYPE** and **PERSON** as the strongest factor groups in the conditioning of pronominal subjects. At level two we have **DISCOURSE CONTINUITY** and **TAM** which show a stronger effect than level three but it is much weaker than level one. Finally, at level three, i.e., the weakest level, we find **POLARITY** and presence of a **MODAL**.

So, the three strongest factor groups to condition pronominal expression are **VERB CLASS**, **CLAUSE TYPE**, and **PERSON**, respectively. Interestingly, these are not the traditional factor groups to strongly condition pronominal expression, at least, not in this particular

order of magnitude of effect. Firstly, **VERB CLASS** does factor as the strongest effect in some studies (Silva-Corvalán, 2001; Torres Cacoullos & Travis, 2010; Travis, 2005), while **PERSON** is the strongest factor in others (Lira, 1982; Monteiro, 1994b; Otheguy, et al., 2007); however, **CLAUSE TYPE** does not normally fare among the factor groups that strongly condition the variable (Duarte, 1993, 2003; Lira, 1982; Monteiro, 1994b; Otheguy, et al., 2007; Silva-Corvalán, 2001; Torres Cacoullos & Travis, 2010; Travis, 2005, 2007).

The second level of effect observed in these data consists of two factor groups, namely **DISCOURSE CONTINUITY** and **TAM**. The former has been consistently found to condition pronominal expression across several dialects of Spanish and in formal letters in BP. In both previous studies and this one, the data points to a pattern of referents that, on the one hand, are repeated tend to be realized with unexpressed referents. On the other hand, referents that are not repeated across sequential clauses tend to be realized pronominally.

TAM has been examined in terms of morphological ambiguity and how it interacts with pronominal expression. Researchers, especially those following a generative framework, have tirelessly claimed that the ambiguity of certain TAMs condition the occurrence of pronominal subjects. Some functional studies, however, claim that morphological ambiguity of TAMs does not tell the entire story, rather, the function of TAMs in discourse are more powerful predictors of the occurrence of pronominal subjects.

After reviewing these results so far it becomes clear that there are other elements impacting the effects of these factor groups in the realization of pronominal subjects. While there is validity in the scientific inquiry of examining these factor groups across the three persons, their combined examination is clearly overshadowing the real effects of each of these factor groups.

In the next three chapters I will move from this general mode of inquiry and attempt to understand the more complex intricacies that are governing how each of the persons interact with pronominal expression. This is going to be achieved in three parts, (a) each person has been submitted to a VRAs to test the effect of each of the factor groups discussed in this section, then (b) I will examine the data further by looking at the way the frequency of certain verbs with each of the persons affect the way expression is realized for that particular person, and finally, (c) I will explore the role of constructions of subjects and predicates and the role of the environments that were excluded from the analyses in the realization of pronominal subjects.

5 RESULTS OF SEPARATE VARIABLE RULE ANALYSES

This chapter presents the results for separate VRAs conducted on each of the three persons that are being examined in this study. The factor groups, or independent variables, that have been included in the analyses are the same ones used to conduct the general analysis described in Chapter 4, namely **TAM**, **VERB CLASS**, **CLAUSE TYPE**, **MORPHOLOGICAL IRREGULARITY**, presence of **MODAL**, **POLARITY**, and **DISCOURSE CONTINUITY.** As I am considering the different persons independently, **PERSON** is not included in the analysis. Thus, a total of three analyses are presented in this chapter, and each one of the analyses containing all predictors were successful in predicting the environments most likely for expressed pronouns to be realized. Each model established a number of different independent variables as statistically significant in promoting the occurrence of pronominal subjects. This finding underscores the significance of pursuing this type of analysis for each person, and each is behaving significantly different than the other two.

5.1 Introduction

In order to understand the way each person is behaving in relation to pronominal subjects, I used the comparative variationist method (Poplack & Tagliamonte, 2001; Torres Cacoullos & Aaron, 2003; Torres Cacoullos & Travis, 2010) to test the hypotheses established in Chapter 3. This method will allow us to draw comparison between the three persons and by doing so, we will be able to observe the parallelism in the structure of subject expression across the three persons. This method also allows us to observe how expression behaves across each person with the same set of constraints.

5.2 1sg Subjects

Results for 1sg subjects are presented in Table 17. A total of 3,447 predicates occurring with 1sg subjects were examined in this analysis (rate of expression = 65%).

Parallel to the overall results, 1sg expressed subjects are strongly conditioned by **VERB CLASS** and **CLAUSE TYPE**, and these two factor groups are close to three times as strong as the next strongest conditioning factor, **DISCOURSE CONTINUITY.** Also significant, though having a weak effect, are **MORPHOLOGICAL IRREGULARITY** and **POLARITY**. **TAM** and **PRESENCE OF MODAL** were not chosen by this model as significant factors in the conditioning of pronominal expression. It seems that, compared to the other factor groups, the distribution of the latter two is likely due to chance, having no effect on the distribution of pronominal 1sg subjects.

In the following subsections I will describe the results for each of the factor groups selected as significant on the conditioning of pronominal 1sg subjects.

Table 16. Multivariate Rule Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 1sg subjects.

Total *N*	3447			
% *expressed*	64.7			
Corrected *Mean*	.671			
	Probability	**% expressed**	**N**	**% data**
Verb class				
Speech	.73	84.3	523	15.2
Possession	.62	74.6	339	09.8
Other	.51	67.4	1225	35.5
Relational	.48	61.5	239	06.9
Perception	.46	59.0	122	03.5
Cognition	.34	49.1	999	29.0
Range 39				
Clause type				
Subordinate	.76	86.7	528	15.3
Main	.45	60.7	2919	84.7
Range 31				
Discourse continuity				
Diff Subj	.54	68.0	1923	55.8
Same Subj & Diff TAM	.50	66.7	684	19.8
Same Subj & Same TAM	.41	55.6	840	24.4
Range 13				
Morphological irregularity				
Regular	.54	66.2	1649	47.8
Irregular	.47	63.3	1798	52.2
Range 07				
Polarity				
Affirmative	.51	66.4	2953	85.7
Negative	.45	54.5	494	14.3
Range 06				
TAM				
Imperfect	[.54]	72.9	487	15.0
Present	[.50]	66.1	1884	58.1
Preterit	[.49]	61.5	871	26.9
Range *n.s.*				
Modal				
Present	[.53]	73.7	495	14.4
Absent	[.50]	63.2	2952	85.6
Range *n.s.*				

Total Chi-square = 605.2260; Chi-square/cell = 1.3913; Log likelihood = -2034.626

5.2.1 *Verb Class*

VERB CLASS is the strongest predictor to the realization of pronominal subjects in these data. Table 17 reproduces the results from Table 16 for this factor group. **SPEECH** and **POSSESSION** predicates largely favor the realization of pronominal subjects while **COGNITION** predicates strongly disfavor the realization of these subjects. The factors **'OTHER'** slightly favors pronominal subjects and '**RELATIONAL**' and '**PERCEPTION**' slightly disfavor them.

Table 17. Result for verb class from VRA for 1sg subjects.

Total *N*	3447			
% *expressed*	64.7			
Corrected *Mean*	.671			
	Probability	**% expressed**	**N**	**% data**
Verb class				
Speech	.73	84.3	523	15.2
Possession	.62	74.6	339	09.8
Other	.51	67.4	1225	35.5
Relational	.48	61.5	239	06.9
Perception	.46	59.0	122	03.5
Cognition	.34	49.1	999	29.0
Range 39				

It must be noted that **SPEECH** and **COGNITION** verbs consist of 44% of the data, nearly half, which suggests that each class of verbs has a specific role in the way expression patterns with 1sg subjects.

SPEECH verbs account for 15% of all verbs occurring with 1sg subjects. Within this class, the verb *dizer* 'to say' alone accounts for 57% (N = 290/523) of all **SPEECH** predicates as can be seen in Figure 1. Also, the verb *dizer* 'to say' shows a very high rate of expression, 87% (N = 252/290), significantly higher than the overall rate for 1sg of 65%.

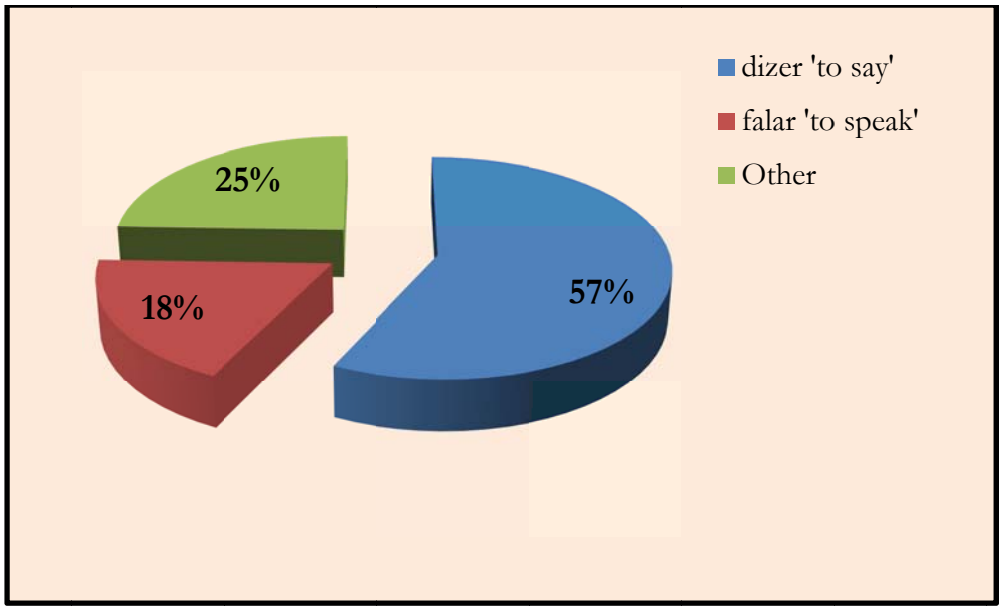

Figure 1. Distribution of speech predicates with 1sg subjects.

The second highest frequent verb in this class is *falar* (N = 98/523) and it also shows a high rate of expression, 83% (N = 81/98).

The '**OTHER**' speech verbs consists of verbs that do not have high enough frequency to stand out as the other two, but similarly to the 2 highest frequent members, this group of verbs shows a high tendency to occur with pronominal 1sg subjects, 80% (N = 108/135). The rates of expression for each of the verbs can be seen in Figure 2

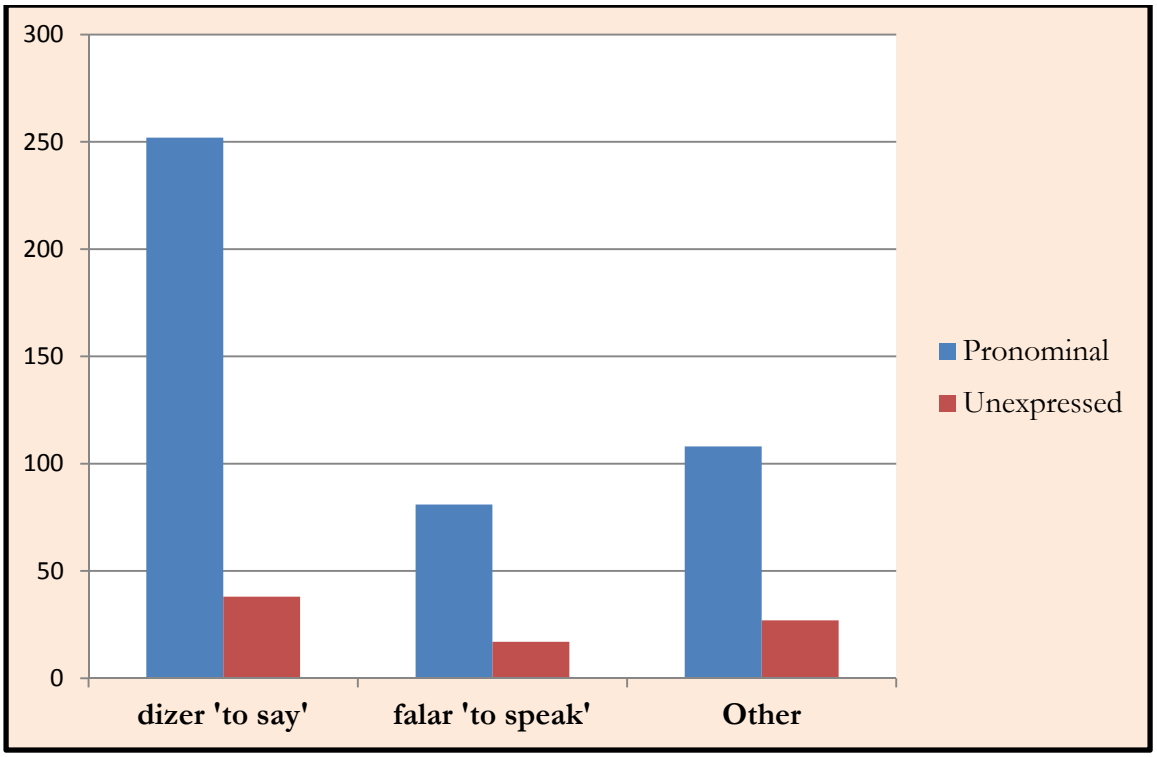

Figure 2. Subject realization in speech predicates (N = 523).

Thus, **SPEECH** predicates with 1sg subjects act as a class, uniformly favoring expressed subjects, as illustrated in (39) and (40).

(39) *Eu digo, varia de governo pra governo.*
'I say, it varies with government.'
(Inq. 7:1243)

(40) *Eu te falei, eu peguei os telefones de vários técnicos de fora.*
'I told you, I got the phone number of several technicians.'
(Inq. 34:39)

While **SPEECH** predicates are highly frequent with 1sg subjects and strongly favor pronominal expression, **COGNITION** verbs are similarly frequent with 1sg subjects yet they disfavor pronominal expression. These predicates account for 29% of all predicates occurring with 1sg subjects. In this verb class, 73% of these predicates consist of two verbs, namely *achar* and *saber* as is evidenced in Figure 3.

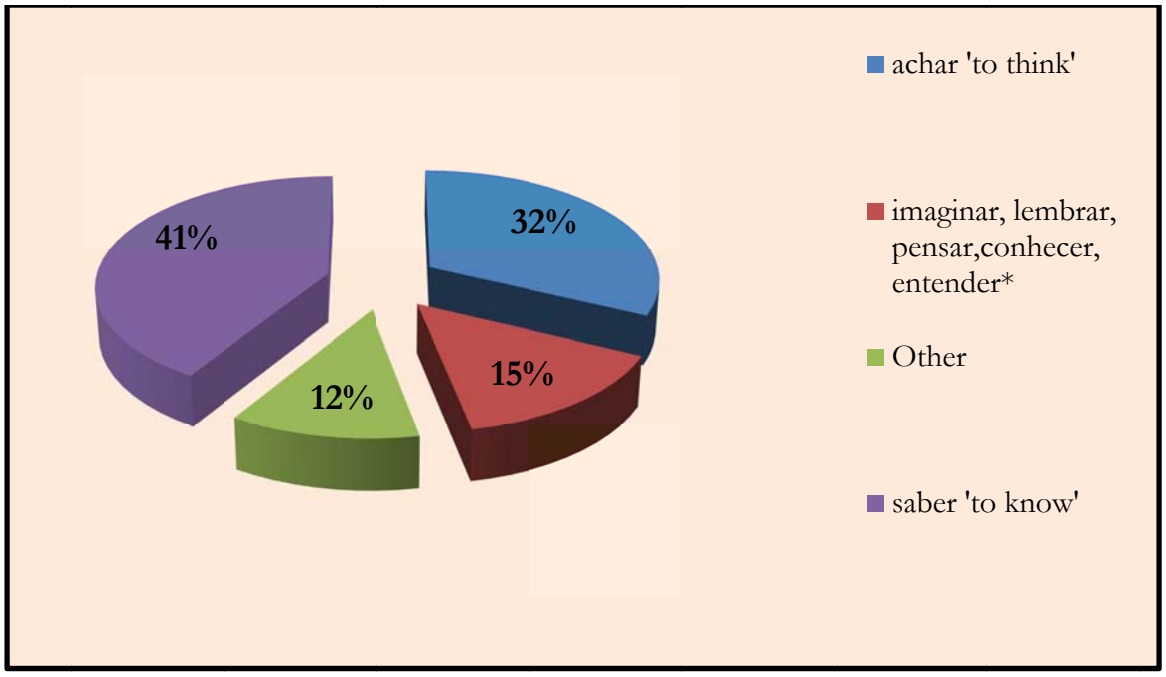

Figure 3. Cognition predicates that co-occur with 1sg subjects.
* These four verbs represent the entirety of this group, they are 'to imagine', 'to remember', 'to think', 'to know', and 'to understand' respectively.

It should be noted that the effect observed in the class of **COGNITION** predicates is detected throughout the entire class with the exception of one member, namely *saber* 'to know' which demonstrates an opposite effect from the one observed in the other members of this class, illustrated in (41) and (42). The fact that **COGNITION** predicates disfavor pronominal subjects is that mostly attributable to the entire class, with the exception of one verb, that disfavors expression as a whole. These patterns are illustrated in Table 18.

(41) *A:* *Depois eu tenho também dicionário da Bíblia... que até um... um amigo meu o pastor S. de Cuba que me deu... aquele... que eu entrevistei*
B: *Sei.*
A: *Que eu fui fazer pesquisa.*
'A: Besides I also have the Bible dictionary ... which ... a friend of mine, pastor S. Cuba gave it to me... that one .. who I interviewed
B: (I) know
A: When I was researching'
(C33:732)

(42) *Eu acho que todo mundo deve além da sua língua deve também carregar uma língua estrangeira.*

'I think everyone should learn a foreign language besides their first language.'
(C47:385)

These findings are very promising in support of our hypotheses that the pattern of pronominal expression is not manipulated across the various lexical items in one single way, but rather, it is locally defined by each lexical item, in this case each construction of person and predicate, and the combination of such patterning compose the more general syntactic pattern we call pronominal expression.

Table 18. Distribution of cognition predicates according to their rates of 1sg pronominal expression.

	Pronominal
achar 'to think'	118 36.7%
imaginar 'to imagine', *lembrar* 'to remember', *pensar* 'to think', *conhecer* 'to know', *entender* 'to understand'	53 34%
Other	31 28.7%
saber 'to know'	293 71.1%
Total	493 49%

5.2.2 Clause type

CLAUSE TYPE is the second strongest factor group conditioning 1sg pronominal subjects in these data. The hierarchy of constraints observed in Table 16, reproduced here as Table 19, illustrate that **SUBORDINATE** clauses strongly favor the use of pronominal subjects and **MAIN** clauses slightly disfavor them.

Table 19. Hierarchy of constraints for clause type in the VRA for the conditioning of 1sg pronominal subjects.

Total *N*	3447			
% *expressed*	64.7			
Corrected *Mean*	.671			
	Probability	**% expressed**	**N**	**% data**
Clause type				
Subordinate	.76	86.7	528	15.3
Main	.45	60.7	2919	84.7
Range 31				

This finding is very surprising because (a) it goes directly against the direction that the pattern of expression is expected to take, and (b) it goes directly against the behavior of these clauses in situation of change in progress.

Concerning the direction that the pattern of expression is expected to take, generative accounts of subject expression have widely argued that **SUBORDINATE** clauses are the loci for unexpressed mentions to be realized because of the very nature of **SUBORDINATE** clauses to hold old referents as their arguments. In this case, then, it is expected that **SUBORDINATE** clauses would favor unexpressed subjects. While it is agreed that **SUBORDINATE** clauses do offer a locus for the occurrence of unexpressed subjects, such a hypothesis can only be raised for headed-relative clauses whose subject is the same as the one in the matrix, or main clause. In other types of **SUBORDINATE** clauses and non-headed-relative clauses the hypothesis does not hold completely as can be seen in these data and has also been attested in other studies (Duarte p.119).

Concerning the behavior of these clauses in situations of change, it has been suggested that **SUBORDINATE** clauses tend to retain older syntactic patterns in a language, thus making them less vulnerable to new patterns that emerge latter on (Bybee, 2002a; Deutscher, 2000). Following this argument, thus, it can be inferred that **MAIN** clauses are

more innovative and susceptible to being used with newer patterns in a language. In short, these premises are in disagreement with the findings observed in these data.

While the motivation for the patterns documented in this section are still unknown, I would like to propose that the effect of **SUBORDINATE** clauses is, in part, conjoined with **TAM** in favoring the realization of pronominal subjects. Looking at the distribution illustrated in Table 20, it can be seen that the effect of **SUBORDINATE** clause is very strong with the **PRETERIT** and **IMPERFECT** TAMs than with the **PRESENT**[32].

Table 20. 1sg subject realization according to clause type and TAM.

		Main		Subordinate		Total	
		N	*%*	*N*	*%*	*N*	*%*
Imperfect							
	Pronominal	235	66.2	120	91	355	73
	Unexpressed	120	33.8	12	9	132	27
Preterit							
	Pronominal	464	62	112	93	576	66
	Unexpressed	286	38	9	7	295	34
Present							
	Pronominal	976	59	182	81	1158	61
	Unexpressed	684	41	42	19	744	40
Total		2765	85	477	15	3260[33]	

The past TAMs in **SUBORDINATE** clauses show rates of pronominal expression of over 90%, while the present is 10 percentage points below these (at 81% expression). In **MAIN** clauses, we can observe a similar pattern, even though it is weaker than the one observed in **SUBORDINATE** clauses, with the past TAMs, more strongly for the **IMPERFECT**, showing rates of expression above the sixtieth percentile while the present is lagging behind.

[32] While the present still shows strong favoring for pronominal subjects in subordinate clauses, it does not achieve the same magnitude of effect that can be observed for the past TAMs.

[33] This total does not include the tokens for the future TAM which was not included in the analysis for TAM.

Hence, what we are really seeing here is not so much an effect of **CLAUSE TYPE** on pronominal expression, but of **CLAUSE TYPE** and **TAM** combined on pronominal expression. Table 20 really permits us to see that the past TAMs are at play in conditioning the realization of pronominal subjects in **SUBORDINATE** clauses; despite **TAM** having not been selected as significant for this particular VRA it is playing a secondary part of the conditioning of the variable.

5.2.3 Discourse Continuity

The VRA for 1sg subjects revealed that **DISCOURSE CONTINUITY** plays a major role in the conditioning of variable subject expression. This has already been reported as a factor group to have a strong effect in the realization of expressed subjects in the Spanish of Los Angeles (Silva-Corvalán, 1994), Caracas (Bentivoglio, 1983), Puerto Rico (Ávila-Shah, 2000), Colombia (Travis, 2005), and New Mexico (Torres Cacoullos & Travis, 2010; Travis, 2007). Similar findings have been reported for BP, specifically for the dialect of Rio de Janeiro in Lira (1982) and Paredes Silva (1993, 2003). The results shown here agree with these previous studies both in the magnitude of effect and the hierarchy of constraints as can be seen in Table 21.

Table 21. Hierarchy of constraints for discourse continuity in the VRA for the conditioning of 1sg pronominal subjects.

Total *N*	3447			
% *expressed*	64.7			
Corrected *Mean*	.671			
	Probability	**% expressed**	**N**	**% data**
Discourse continuity				
Diff Subj	.54	68.0	1923	55.8
Same Subj & Diff TAM	.50	66.7	684	19.8
Same Subj & Same TAM	.41	55.6	840	24.4
Range 13				

Paredes Silva (1993, p. 43) raises the argument that the effect observed here is not necessarily an artifact of the switch in reference, but indeed a change of discourse topic. She elaborated a detailed layered system to account for not only changes of reference and tense, but also changes of "topic chain"[34] in that not only the referent and the tense change, but also the event that is being described and the narrative sequence. She shows that as discourse connectedness decreases, expression increases, and vice-versa, just as has been observed in these data. As the findings show, expressed subjects are more likely to emerge in contexts of less connected discourse as is illustrated in example (43) below. In this example, the first clause *eu gosto* 'I like' and the last one *eu num vou assistir* 'I won't watch' are separated by another clause whose referent is one other than a 1sg subject, namely *sabe* 'you know'.

(43) *Eu gosto muito também de esporte sabe?*
Sobretudo quando é[35] *o Brasil*
Logicamente que eu num vou assistir.
"I like sports a lot, you know?
Especially when Brazil is playing
Logically I won't watch it.

(Inq. 33:1008)

Example (44) below shows the predicted pattern, but in the opposite direction, that is, more continuous subjects tend to be realized with less linguistic form. The predicates in this example are underlined to show the continuity of reference without an overt pronoun. This supports a large body of literature that has reported the effect of this factor group in the realization of pronominal subjects. Such finding leads us to infer on the universality of discourse continuity in natural languages (Chafe, 1994; Givón, 1983a; Levinson, 1987).

(44) *Doc. - certo... agora ingredientes assim de uma comida que a senhora gosta muito a senhora conhece?*

[34] See Li and Thompson for a discussion of "topic chains" (Li & Thompson, 1976).

[35] This predicate did not factor in the count because it is a non-referential.

Inf. - *conheço...*
mas não tenho disposição nem mesmo pra fazer isso aí...
esses pratos deliciosíssimo... /pesar de conhecer os ingredientes...
tenho toda a receita
mas não tenho... aptidão
para:: se habilita::r a fazer isso
"Doc. – right…now the ingredients of a dish that you like it very much, you know it?
Inf. – (I) know it…
but (I) am not willing to make even those…
delicious dishes… even though (I) know all the ingredients…
(I) have the entire recipe
but (I) am not good at it
to be able to do these things"

(Inq. 09:171-178)

These results show that it is not continuity of subject alone, but also continuity of TAM. Note the weights between the factors same subject & same TAM (.41) and same subject & different TAM (.50), just with a shift of TAM we see a 22% increase, while from the latter to a difference in subject the weights show an increase of only 8%. So, what they show is that coreferentiality is not enough to condition the realization of pronominal subjects, as can be seen also in (45) below. In this example, we have a string of clauses with coreferential subjects, but some are expressed and some are not. Those that are unexpressed are those where the TAM is the same as the preceding clause; those that are expressed are those where there is a change of TAM, e.g. from the present in clause 5 to the preterit in clause 6, and then to the imperfect in clause 7.

(45) *Speaker B:* *eu dei$_1$ aula no Estado... Colégio Justiniano de Serpa*
Speaker A: *colégio do Estado*
Speaker B: *Serpa entrei$_2$ em cinqüenta e oito*
Speaker A: *Justiniano de Serpa*
Speaker B: *fiz$_3$ trinta anos pedi$_4$ minha aposentadoria ... eu estou$_5$ aposentado do segundo grau... agora eu comecei$_6$ no magistério superior na escola de enfermagem ... nessa época eu era$_7$ professor da escola Doméstica.*
"Speaker B: I taught-pret at a school in the state

Speaker A: state school
Speaker B: Serpa I got in-pret in fifty-eight
Speaker A: Justiniano de Serpa
Speaker B: (I) turned-pret thirty years old (I) requested-pret my retirement ... I am-pres retired from high school... now I began-pret to teach university level classes at the nursing school ... at this time I was-impf a professor of Economics."

(Inq. 47:39)

The implications of the findings in the priming literature (Cameron, 1994; Cameron & Flores-Ferrán, 2004; Hochberg, 1986; Pickering, Branigan, Cleland, & Stewart, 2000) reveal very powerful results when TAM of main verb is cross tabulated with discourse continuity. As seen in Table 22, when occurring with verbs in the present, discourse continuity has little effect: 1sg subjects are expressed at a rate between 60% and 63% for the different degrees of discourse continuity. For the preterit, however, we observe a much steeper curve of increase in rates of expression from 46% in contexts of maximum continuity to 76% in contexts of minimum continuity, with a steady rise followed throughout. The imperfect, on the other hand, shows a similar rise from 54% to 84%, but without the steady increase across the different degrees of continuity that we see in the preterit. We do not see, however, a clear change in pronominal expression when the referents change.

Table 22. Rates of 1sg pronominal realization according to discourse continuity and TAM.

	Present	**Preterit**	**Imperfect**
Same Subj & Same TAM	313 59.3%	95 46.3%	45 54.2%
Same Subj & Diff TAM	172 62.8%	130 66.7%	112 74.7%
Diff Subj	673 62.2%	351 73.5%	198 67.4%
χ^2	.553	.000*	.000*

* Significant at p<.005

These results reveal that discourse continuity is not an overarching factor that applies across all tenses in these data. Rather, each TAM patterns differently. Moreover, the way discourse is organized does not fully explain the nature of the phenomenon. What seems to be clear, however, is that there seems to be a convergence between cognitive processes mingled with online language processing and the need to establish communication.

5.2.4 *Morphological Irregularity*

Morphological irregularity measured the effect of the regularity of inflectional morphology on pronominal expression. It is the fourth strongest factor group in the conditioning of pronominal 1sg subjects in these data. However, it must be noted that its effect compared to the three strongest factor groups is weak (with a Range of just 7, half that of the next strongest factor group, Discourse Continuity) and should be taken with caution. The hierarchy of constraints is reproduced in Table 23 below. The findings suggest that regular verbs tend to favor pronominal subjects and irregular ones tend to disfavor the occurrence of pronominal subjects.

Table 23. Hierarchy of constraints for morphological irregularity in the VRA for the conditioning of 1sg pronominal subjects.

Total *N*	3447			
% *expressed*	64.7			
Corrected *Mean*	.671			
	Probability	**% expressed**	**N**	**% data**
Morphological irregularity				
Regular	.54	66.2	1649	47.8
Irregular	.47	63.3	1798	52.2
Range 07				

The hypothesis established by this factor group is borne out in that neutralization of morphological markings condition the occurrence of pronominal subjects, and the lack

thereof discourages the use of these pronouns. This is second device to test for the hypothesis of ambiguity, here not just of TAM, but of the verb declension as well.

However, while these results are suggestive of the effect of verbal morphology on the occurrence of pronominal subjects, I will show later in this work that this is, for the most part, an effect of certain high frequency verbs, which in turn tend to be highly irregular (Bybee, 1985).

5.2.5 Polarity

Polarity is the weakest factor group in conditioning the realization of pronominal 1sg subjects in these data. The results in the VRA are reproduced here in Table 24 below. It can be seen that affirmative statements favor the realization of pronominal 1sg subjects while negative statements disfavor the occurrence of such pronouns. This pattern has already been noted in the literature by Duarte (1993, 2000, 2003). Although she did not offer an explanation for this patterning, I attribute it to the fact that the presence of a negative particle induces the non-realization of 1sg subjects because these statements are of less frequency, yet are much more marked syntactically, which renders them more entrenched in their composition, leading to a much tighter constituency between elements, leaving the opportunity for expressed subjects to emerge very marginally. In other words, negative markers and predicates have, over time, formed constructions without pronominal subjects that are somewhat impervious to the more frequent pattern of expression.

Table 24. Hierarchy of constraints for polarity in the VRA for the conditioning of 1sg pronominal subjects.

Total *N*	3447			
% *expressed*	64.7			
Corrected *Mean*	.671			
	Probability	**% expressed**	**N**	**% data**
Polarity				
Affirmative	.51	66.4	2953	85.7
Negative	.45	54.5	494	14.3
Range 06				

Another possible explanation for the conditioning of pronominal subjects by polarity will be explored in Chapter 7, where I will argue that there may be a construction effect in that subjects, negation markers, and predicates form a constituent that may or may not be realized with expressed subjects, and this patterning may determine why this factor group shows such significance in the conditioning of the variable for 1sg subjects.

5.2.6 Summary

In this section I presented the results for the VRA for the conditioning of 1sg subject expression and observed that five of the seven predictors included in the analysis were selected as significant in predicting the occurrence of 1sg pronominal subjects, namely verb class, clause type, discourse continuity, morphological irregularity and polarity.

For verb class it was noted that speech and possession predicates strongly favor expression, whereas cognition predicates strongly disfavor expression. The other factors showed a more or less neutral effect. These effects were attributed to lexical effects in each of the classes of predicates, viz. *dizer* 'to say', *achar* 'to think', *saber* 'to know', and *ter* 'to have'. These verbs, it is argued, contributed strongly to the patterns of pronominal expression, or lack thereof, in these data.

Clause type showed very surprising results in that subordinate clauses were the ones to favor pronominal expression the most. However, it was demonstrated that these effects were not produced by clause type alone, but rather by an interplay between subordinate clauses and the past TAMs, which are nearly categorical in their favoring for 1sg pronominal subjects.

Discourse continuity was the third strongest factor to condition 1sg subject expression, and again it was a factor group that was deeply intertwined with TAM. It was shown that discourse continuity alone does not paint the overall picture, but an understanding of how it works along with the different TAMs provided us with a great insight on the behavior of 1sg subject expression. I remarked on the fact that each TAM shows its own pattern with discourse continuity, viz. the present is fairly neutral in that it does not show much increase in pronominal expression across the different levels of continuity; the preterit, on the other hand, demonstrated an almost classical, hypothesis-fulfilling pattern of logarithmic increase across each of the four different levels; the imperfect, on the other hand, presented itself as the TAM that most affects the continuity scale in that a change a of TAM would be more predictive of pronominal subjects than the change of referent as can be assumed to be the pattern for the other TAMs.

Morphological irregularity and polarity were the weakest factor groups in the conditioning of 1sg pronominal subjects in these data. We observed that affirmative statements and regular verbs slightly favor pronominal subjects. Alternatively, negative statements and irregular verbs slightly disfavor the use of pronominal 1sg subjects in this corpus. For both factor groups, a frequency effect was raised as the possible explanation for

these patternings. Irregular verbs are frequent in languages and have been prolifically used with unexpressed referents, which solidified their occurrence without pronominal subjects.

The most striking observation to emerge from this data relate specifically to the need for careful scrutiny of the data beyond the VRAs. While VRAs are resourceful and depict a general idea of what the data is doing and how it is behaving, it does not paint the entire picture. As I hope to have begun to demonstrate in this chapter, it is necessary to transcend the results presented in a particular VRA to understand the real nature of the phenomenon at hand. In our case, the analysis of 1sg subjects has showed that there are indeed local patterns of subjects, predicates, and TAMs acting upon larger syntactic patterns in the language, more specifically, subject expression. These patterns will be explored more fully in Chapters 6 and 7. Now I will turn to discuss of the VRA conducted on the 2sg subjects.

5.3 2sg Subjects

Results for the VRA on the conditioning of 2sg subject expression are presented in Table 25 below. A total of 1,689 predicates with 2sg subjects were subjected to the analysis. The rate of expression for 2sg subjects is 53.4%, which is much lower than the rate observed for 1sg subjects of 64.7%. This in itself is suggestive that the two datasets may behave differently from one another and should thus be analyzed separately.

The overall model presented in Table 25 presents five predictors, or factor groups, as significant in the conditioning of 2sg pronominal subjects, namely **VERB CLASS**, **CLAUSE TYPE**, **TAM**, **MORPHOLOGICAL IRREGULARITY**, and **MODAL**. In this analysis **DISCOURSE CONTINUITY** and **POLARITY**, which were found to be significant for 1sg, did not fare as significant factors in predicting the occurrence of 2sg pronominal subjects.

Table 25. Multivariate Rule Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 2sg subjects

Total *N*	1689				
% *expressed*	53.4				
Corrected *Mean*	.547				
		Probability	**% expressed**	**N**	**% data**
Verb class					
Possession		.78	82.8	180	10.7
Relational		.70	76.6	94	05.6
Other		.70	73.9	544	32.2
Speech		.48	55.8	86	05.1
Perception		.34	36.7	281	16.6
Cognition		.24	25.4	504	29.8
Range	54				
Clause type					
Subordinate		.79	83.7	246	14.6
Main		.45	48.2	1443	85.4
Range	34				
TAM					
Imperfect		.61	70.2	114	07.3
Present		.56	53.2	1220	78.4
Preterit		.48	48.9	222	14.3
Range	13				
Morphological irregularity					
Irregular		.56	58.8	808	47.8
Regular		.45	48.5	881	52.2
Range	11				
Modal					
Present		.58	61.7	311	18.4
Absent		.48	51.5	1378	81.6
Range	10				
Discourse continuity					
Same Subj & Same TAM		[.52]	61.4	308	18.2
Same Subj & Diff TAM		[.50]	56.8	229	13.6
Diff Subj		[.47]	50.6	1152	68.2
Range	*n.s.*				
Polarity					
Affirmative		[.50]	53.4	1551	91.8
Negative		[.47]	52.9	138	08.2
Range	*n.s.*				

The way the factor groups are divided in terms of their magnitude of effect, or range, suggests two levels of strength of conditioning, that is, level one with **VERB CLASS** and **CLAUSE TYPE** with very strong ranges, and the latter is almost three times stronger than the next significant factor group. Thus, I will argue here that these two predictors are much better predictors of 2sg pronominal subjects than the remaining factor groups. The second level is much weaker and it consists of **TAM**, **MORPHOLOGICAL IRREGULARITY**, and **MODAL**. These predictors, while significant, are not powerful predictors of the realization of 2sg pronominal subjects, at least not as strong as the two factors present in level one. In the following subsections I will discuss each of the significant predictors and how they affect 2sg pronominal subjects in these data.

5.3.1 Verb Class

VERB CLASS is the strongest factor group in conditioning 2sg pronominal subjects. It is so strong compared to the other factor groups in this analysis that it is one and half stronger than the next strongest factor group, **CLAUSE TYPE**, and a bit over four times stronger than the third strongest factor group, **TAM**. Thus, this finding suggests that these classes of verbs are very reliable predictors of 2sg subject expression. The hierarchy of constraints observed in Table 25 is reproduced below in Table 26. **VERB CLASS** has been found to affect pronominal subjects over and over throughout BP and dialects of Spanish (Lira, 1982; Otheguy, et al., 2007; Silva-Corvalán, 2001; Silveira, 2008; Torres Cacoullos & Travis, 2010; Travis, 2005, 2007). While these studies do not focus on 2sg subjects per se, the effects of verb class are across the board. However, vis-à-vis 1sg subjects, **COGNITION** predicates strongly disfavor the realization of pronominal 2sg subjects. This is in stark contrast to these previous works.

Table 26. Hierarchy of constraints for verb class in the VRA for the conditioning of 2sg pronominal subjects.

Total *N*	1689			
% *expressed*	53.4			
Corrected *Mean*	.547			
	Probability	**% expressed**	**N**	**% data**
Verb class				
Possession	.78	82.8	180	10.7
Relational	.70	76.6	94	05.6
Other	.70	73.9	544	32.2
Speech	.48	55.8	86	05.1
Perception	.34	36.7	281	16.6
Cognition	.24	25.4	504	29.8
Range 54				

In these data, **POSSESSION**, **RELATIONAL** and '**OTHER'** predicates strongly favor the occurrence of pronominal 2sg subjects. On the other hand, **PERCEPTION** and **COGNITION** predicates strongly disfavor the use of pronominal subjects with this person. While **SPEECH** predicates appear to disfavor pronominal 2sg subjects when compared to **POSSESSION**, **RELATIONAL**, and '**OTHER'**, they do not achieve the same effect that **PERCEPTION** and **COGNITION** do. Thus, **SPEECH** predicate will be treated as having a neutral effect henceforth. And it is worth noting that this effect differs from 1sg subjects, where these predicates most favored expression.

POSSESSION predicates account for nearly 11% of all 2sg subjects and it is the fourth most frequent verb class to occur with these subjects. The most frequent member of this class is the verb *ter* 'to have', which alone accounts for 73% of all predicates in this class, of which 87% are realized with pronominal 2sg subjects as can be seen in Figure 4.

RELATIONAL predicates are the second strongest verb class in favoring 2sg pronominal subjects. The most frequent verb in this class is *ser* 'to be' (N=56/94, or 60%)

which highly favors pronominal subjects (88% or 49/56). I attribute the effect of this class to the strong patterning of pronominal 2sg subjects and the verb *ser* 'to be'.

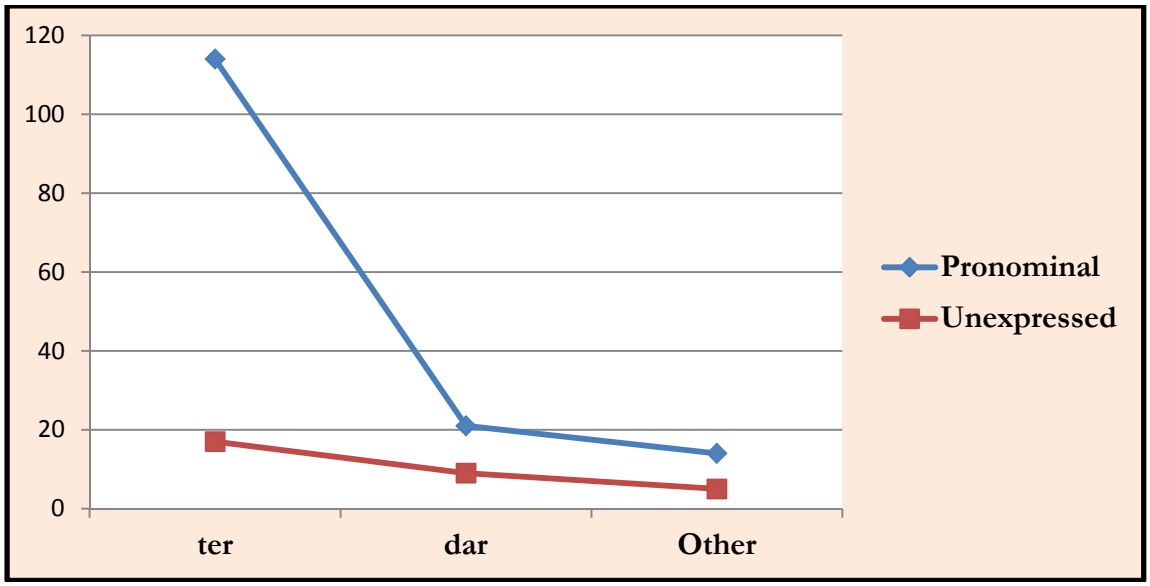

Figure 4. Distribution of possession predicates with 2sg subjects according to their rates of pronominal expression.
$\chi^2 = 6.194$; $p = .045$

The third strongest factor is the category 'other', which is as strong as **RELATIONAL** based on their weights. This factor harbors two very frequent verbs, namely *fazer* 'to do/make' and *ir* 'to go'. These two verbs account for 17% (N=90) of the entire factor group, showing rates of expression of 81% and 92% respectively. This is strong evidence that these verbs are playing a much more active role in determining the occurrence of 2sg pronominal subjects. It must be noted, however, that while these two verbs are clearly affecting the realization of the variable, the entire category of '**OTHER'** predicates favors the occurrence of pronominal subjects. In short, the general pattern that verbs are following is that of pronominal expression.

Predicates of **PERCEPTION** and **COGNITION** disfavor the expressed subjects, with the latter strongly disfavoring it. With verbs of **COGNITION**, the effect can be clearly observed

with two verbs, namely *entender* 'to understand' and *saber* 'to know'. The two verbs account for 76% of all tokens of **COGNITION** verbs with 2sg subjects (N/504) and they show almost identical patterns of pronominal expression with 2sg subjects.

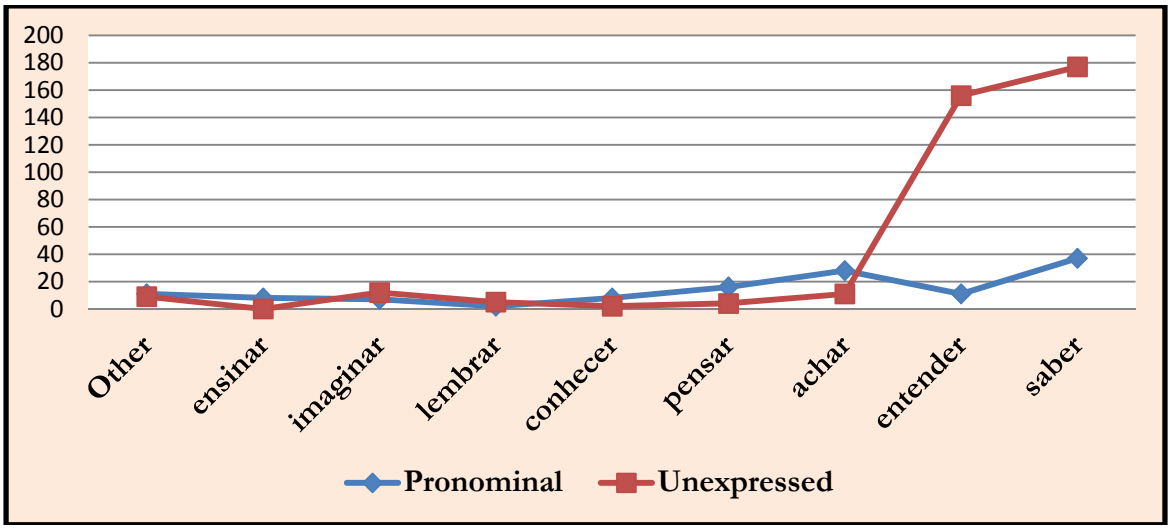

Figure 5. Distribution of cognition predicates with 2sg subjects according to their rates of expression. $\chi^2 = 164.232$; p = .000

As can be seen in Figure 5, 93% (N = 156/167) of all tokens of *entender* with a 2sg subjects occurred with an unexpressed subject. Similarly, the verb *saber* shows a rate of unexpression of 83% (N = 177/214). Thus, it is argued here that these two verbs are strongly responsible for the disfavoring toward the realization of expressed subjects observed in the class as a whole. The remaining verbs, on the other hand, comply with the overall pattern of higher rates of expression. These verbs occur 65% (N = 80/123) of times with pronominal 2sg subjects.

The category of **PERCEPTION** verbs is much less populated than that of **COGNITION** as can be noted in Figure 6 below. It can be seen, though, that one verb in particular seems to be leading the class to the favoring of unexpressed subjects, namely *olhar*, which accounts for 44% of all **PERCEPTION** predicates, of which, 95% are realized with unexpressed subjects.

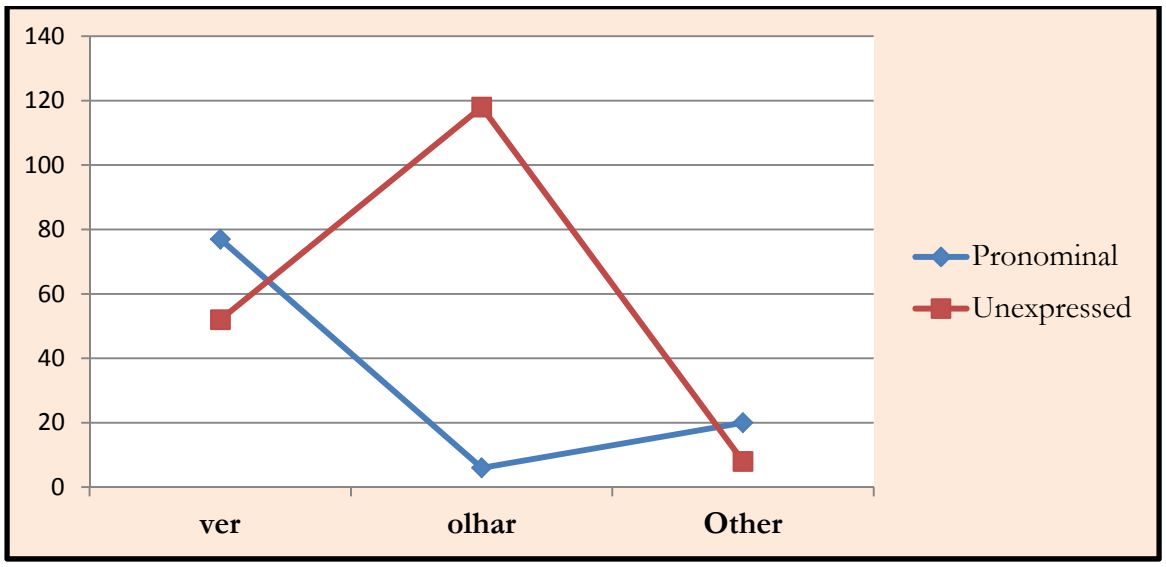

Figure 6. Distribution of perception predicates with 2sg subjects according to their rates of expression. $\chi^2 = 98.121$; p = .000

This section described the effects of **VERB CLASS** on the realization of 2sg pronominal subjects in these data. As was argued for 1sg subjects in the previous section (cf. 5.2.1), this is strong evidence of local, lexical effects at play in the realization of the syntactic pattern of pronominal expression. We can see that the realization of pronominal 2sg subjects is very much determined by the verb that is being used.

A final point worthy of mention about this factor group is that three of these verb classes constitute 57% of the data, namely possession, perception, and cognition, with the latter two disfavoring the use of the variable. This is another piece of evidence to support the claim that these verbs as a class and individually are responsible in determining the seemingly random pattern of pronominal expression in these data. This hypothesis will be pursued further in Chapter 6.

5.3.2 *Clause type*

Clause type is the second strongest factor group in the conditioning 2sg pronominal subjects in these data. The hierarchy of constraints observed in Table 25, reproduced here as Table 27, illustrate that subordinate clauses strongly favor the use of pronominal subjects and main clauses slightly disfavor their use. This also follows the same pattern observed for 1sg subjects.

Table 27. Hierarchy of constraints for clause type in the VRA for the conditioning of 2sg pronominal subjects.

Total *N*	1689			
% *expressed*	53.4			
Corrected *Mean*	.547			
	Probability	**% expressed**	**N**	**% data**
Clause type				
Subordinate	.79	83.7	246	14.6
Main	.45	48.2	1443	85.4
Range 34				

5.3.3 *TAM*

Turning now to the results concerning **TAM**, for analyzing **TAM** instead of the traditional ambiguity. Firstly, the factor group **TAM** captures not only the notion of **AMBIGUITY**, but it also allows us to observe the way discourse is organized. Secondly, I would like to follow Silva-Corvalán in that "es el tiempo verbal, más que la ambigüedad, y aun más específicamente la function de los diferentes tiempos en el discurso, lo que se correlaciona com expresíon del sujeto" (2001, p. 161). In this way, the category **TAM** captures the variation more truthfully than the ambiguity hypothesis. Moreover, a mathematical reason for preferring to use the factor group **TAM** in place of **AMBIGUITY** lies in the fact that initial VRA's showed that ambiguity is not a significant predictor of expressed subjects in general.

The VRA shows that the **IMPERFECT** favors pronominal 2sg subjects (.61), while the **PRETERIT** disfavors their realization (.48). The **PRESENT** also favors the realization of pronominal 2sg subjects (.56), but less than the **IMPERFECT**. Table 28 illustrates the results for this factor group.

Table 28. Hierarchy of constraints for TAM in the VRA for the conditioning of 2sg pronominal subjects.

Total *N*	1689			
% *expressed*	53.4			
Corrected *Mean*	.547			
	Probability	**% expressed**	**N**	**% data**
TAM				
Imperfect	.61	70.2	114	07.3
Present	.56	53.2	1220	78.4
Preterit	.48	48.9	222	14.3
Range 13				

The patterning observed in these data for the **IMPERFECT** suggest that ambiguity may be playing a role in the realization of pronominal 2sg subjects since in this **TAM** all instances are ambiguous since 2sg and 3sg subjects have the same forms in all tenses. This will be discussed in more detail in section 5.5.6. it should also be noted that, though not significant, this is the same direction of effect observed for 1sg subjects.

5.3.4 Morphological Irregularity

The hypothesis tested here was that the degree of regularity of the verb stem would affect the outcome of pronominal expression. It was expected that more **IRREGULAR** verbs would favor unexpressed subjects, and more **REGULAR** verbs would favor expressed subjects, as was found for 1sg. As can be seen in Table 29, these predictions are not borne out. What we find is that **IRREGULAR** verbal forms tend to condition the realization of pronominal subjects

(.56), while the more **REGULAR** verbal forms emerge more often without pronominal subjects (.45).

Table 29. Hierarchy of constraints for morphological irregularity in the VRA for the conditioning of 2sg pronominal subjects.

Total *N*	1689			
% *expressed*	53.4			
Corrected *Mean*	.547			
	Probability	**% expressed**	**N**	**% data**
Morphological irregularity				
Irregular	.56	58.8	808	47.8
Regular	.45	48.5	881	52.2
Range 11				

In order to understand the effects of **MORPHOLOGICAL IRREGULARITY** on pronominal 2sg subjects it is necessary to investigate how different verbs pattern with pronominal subjects. If we look at the way **IRREGULAR** verbs behave in regards to pronominal expression, it can be seen that most verbs in this category favor pronominal expression (the opposite of the expected pattern), except for the verb *saber* 'to know' as documented in Figure 7.

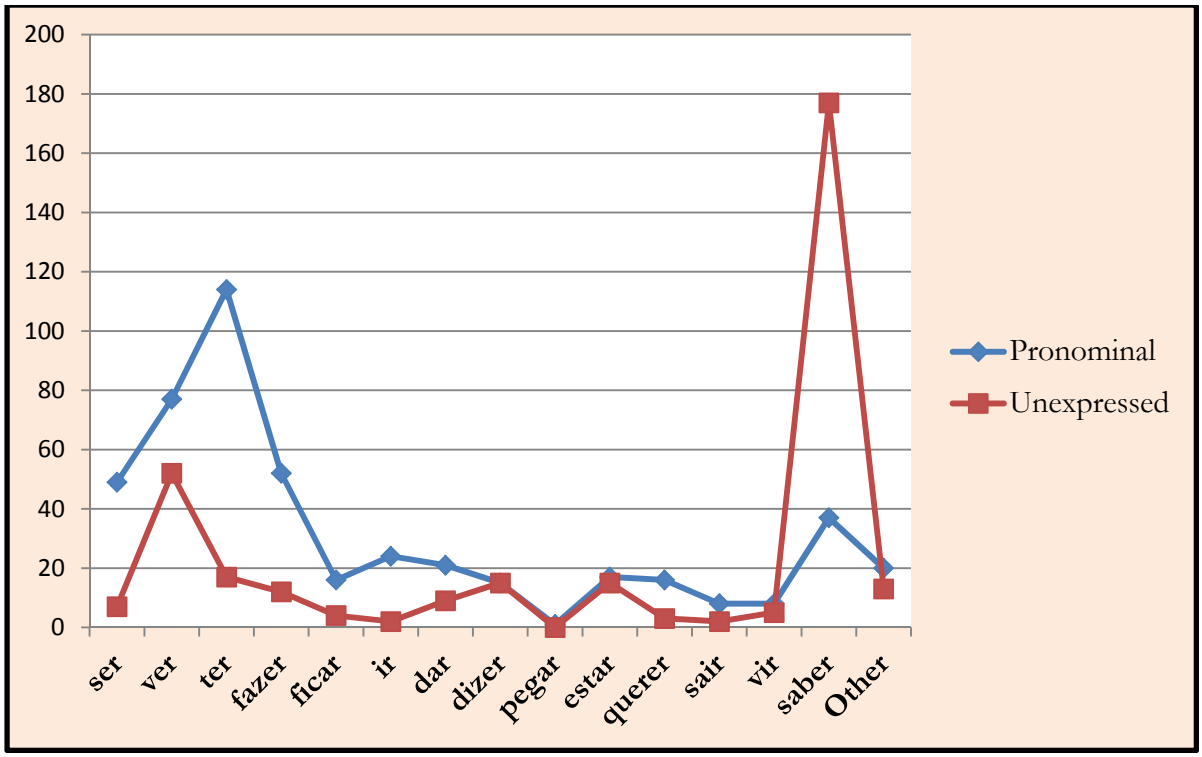

Figure 7. Distribution of irregular verbs according to pronominal expression.
$\chi^2 = 264.062$; $p<.000$
The 'other' category consists of 12 verb types.

Most of **IRREGULAR** verbs favor the expressed 2sg subjects. This finding suggests that lexical patterning with syntactic patterns override morphological ones. In this case, these verbs, Thus, it is imperative to examine them in more detail as we are attempting to do in this work.

REGULAR verbs in these data disfavor the 2sg subject expression. Figure 8 shows the distribution of these verbs with pronominal subjects. The effect of these verbs can be attributed to two verbs in particular, *olhar* 'to look' and *entender* 'to understand'. These two verbs are highly frequent, accounting for 33% (291/881) of the set of regular verbs and they are clearly affecting the way these verbs are patterning with pronominal subjects.

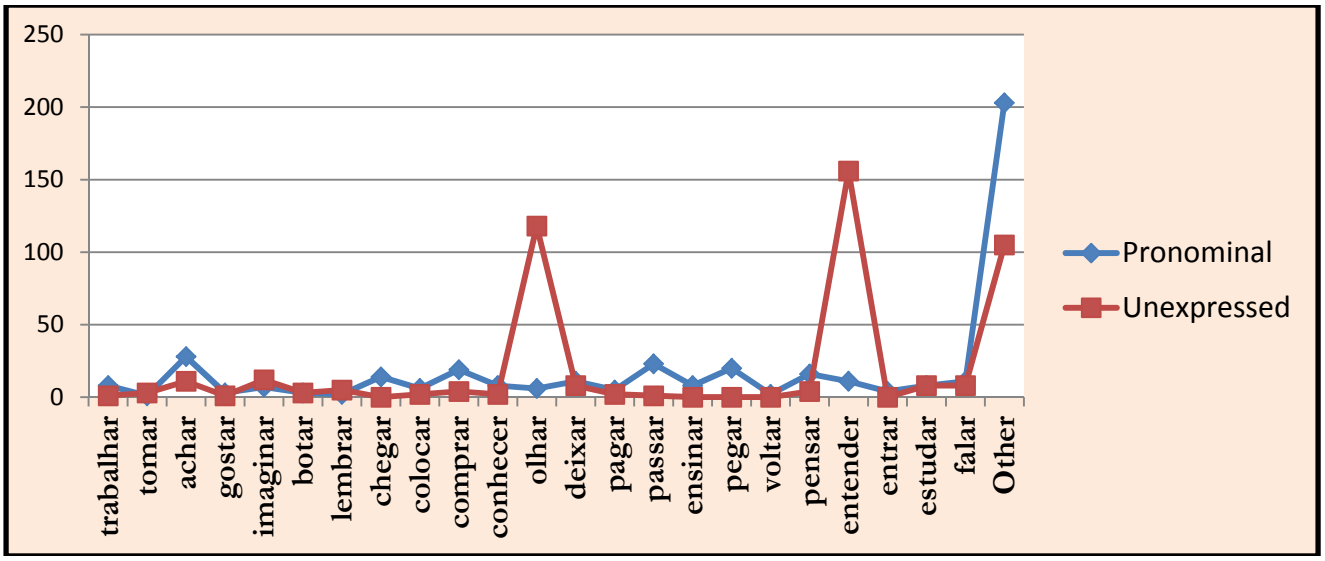

Figure 8. Distribution of regular verbs according to pronominal expression.
$\chi^2 = 368.198$; $p<.000$
The 'other' category consists of 141 verb types

5.3.5 Modal

This factor group was the weakest among the factors selected as significant in the analysis of 2sg pronominal subjects. It can be seen in Table 30 that clauses with a modal are more likely to occur with pronominal 2sg subjects. This finding follows the direction of effect observed for 1sg subjects, even though this was not selected as significant.

Table 30. Hierarchy of constraints for modal in the VRA for the conditioning of 2sg pronominal subjects

Total *N*	1689				
% *expressed*	53.4				
Corrected *Mean*	.547				
		Probability	**% expressed**	**N**	**% data**
Modal					
Present		.58	61.7	311	18.4
Absent		.48	51.5	1378	81.6
	Range 10				

The pattern of pronominal expression with modals can be attributed to three modals in particular, namely *ir* 'to go' (89% expressed, N=75/84), *poder* 'can' (91% expressed, N=41/45), and *ter* 'to have' (84% expressed, N=21/25). These forms account for 49% of all

modals in these data (N=154/311), of which 89% (N=137/154) occur with pronominal subjects. Again, what we can observe is that there are lexical effects at play that affect the conditioning of the variable beyond the predicting power of any statistical analysis.

5.3.6 Summary

The VRA results for the conditioning of 2sg pronominal subjects reveal that these subjects are affected by **VERB CLASS**, **CLAUSE TYPE**, **TAM**, **MORPHOLOGICAL IRREGULARITY** and **MODAL**. These factor groups were selected as significant predictors in the realization of these subject pronouns. After careful examination of each of these factor groups, we have learned that although we have achieved a statistical significant model that accounts for the realization of 2sg subjects, it is fundamental that we go beyond the results illustrated by the analysis and scrutinize the local patterns that make up the syntactic pattern of pronominal expression.

The strongest factor group to condition the realization of 2sg pronominal subjects is **VERB CLASS**. I would venture to say that this is the most important factor group in determining the realization of pronominal subjects. It is observed that **POSSESSION** and **RELATIONAL** predicates strongly favor the conditioning of pronominal subjects. The category **'OTHER'** also favors the realization of pronominal subjects. I claimed that this finding suggests that the pattern of expression is entrenched in the verbs that do not form a class, that is, most verbs in these data. Conversely, **PERCEPTION** and **COGNITION** predicates strongly disfavor the occurrence of pronominal 2sg subjects, with latter disfavoring it the most.

CLAUSE TYPE is the second strongest factor group in conditioning the emergence of pronominal subjects. Just as it was observed for 1sg subjects, **SUBORDINATE** clauses strongly favor the realization of pronominal subjects while main clauses do not. This is most intriguing and it will be discussed further in section 5.5.1.

A third factor group selected as significant was **TAM**, which shows much lower effect than the first two as can be seen by its range (13). The VRA revealed that the **IMPERFECT** and the **PRESENT** favor pronominal subjects, with the former showing a stronger effect, and the **PRETERIT** disfavors them. Despite the fact that these findings suggest an effect of verbal ambiguity, they cannot be taken wholeheartedly because the percentage of the data represented by the **IMPERFECT** TAM is less than 10%, thus granting these findings suggestive at best. Thus, more research needs to be undertaken before the association between the **IMPERFECT** and pronominal 2sg subjects is more clearly understood. Another issue to be examined in future analyses is whether the **PRESENT** and the **PRETERIT** are indeed ambiguous for 2sg subjects.

To assess the effect of verbal regularity on the realization of 2sg pronominal subjects, the factor group morphological irregularity was included in the statistical analysis and it was selected as one of the factor groups to have a significant effect on the occurrence of these subjects. It was observed that **IRREGULAR** verbs favor pronominal subjects and **REGULAR** verbs disfavor them. While these findings are not at all conspicuous, they do reveal a lexical effect in that in both factors, irregular and regular verbs, there is a number of isolated lexical items (e.g. *num sabe* '(you) don't know') that are leading the factor into the pattern they are demonstrating in the statistical analysis.

Finally, the last factor group to have an effect on 2sg pronominal subjects is **MODAL**. In this factor group it was documented that the presence of a modal induces the use of pronominal 2sg subjects. The inverse is applicable, that is, the absence of a modal correlates with the non-use of pronominal 2sg subjects. It was also noted that only three modals (*poder*

'can', *querer* 'want' and *ter que* 'have to') account for nearly fifty percent of all modals documented in this study.

Some of the issues relating to the findings obtained in the VRA for 2sg subjects relate specifically to the claim that has been made across this work that it is fundamental to examine the local patterns that are forming between predicates and subjects to understand the variable and how it plays out in discourse more fully. This was evidenced throughout the several significant factor groups analyzed in this section in that each of them demonstrated to be affected by more local patterns of verbs and subjects. This claim will persist throughout this study.

One last point that needs to be noted relates to discourse continuity. Although it was not selected as significant in this analysis, it goes precisely in opposite direction to what was predicted and to what was observed for 1sg subjects. This can be attributed to (a) the nature of 2sg subjects in discourse in that they are not prone to occur in contexts of maximum continuity, and (b) to lexical effects. In Chapters 6 and 7, I will discuss the effect of frequency and constructions on the realization of pronominal subjects, and I will argue that what can be seen here for 2sg subjects can be explained by these two concepts. Among the three subjects being examined in this study, 2sg subjects are the ones most affected by the proportion of high frequency verbs and constructions that occur in these data. After excluding all high frequency verbs and constructions from this data set, only 23% remains, which tells how important these tokens that were excluded are in the makeup of any syntactic pattern that affects these subjects.

5.4 3sg Subjects

The results for the VRA on the conditioning of 3sg pronominal subjects are documented in Table 31 below. A total of 2,930 predicates were included, with an overall rate of expression of 48%, thus markedly lower than 1sg (rate of 64%) and slightly lower than 2sg (53%)

From the seven predictors included in these separate analyses, only four were selected as significant to the conditioning of 3sg pronominal subjects in these data, namely **MODAL**, **DISCOURSE CONTINUITY**, **CLAUSE TYPE**, and **TAM**. Unlike 1sg and 2sg subjects, 3sg subjects do not appear to be affected by **VERB CLASS**. In the following subsections I will discuss each of the selected factor groups and how they affect 3sg pronominal expression.

Table 31. Multivariate Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 3sg subjects

Total *N*	2930			
% *expressed*	47.7			
Corrected *Mean*	.475			
	Probability	**% expressed**	**N**	**% data**
Modal				
Absent	.53	50.5	2358	80.5
Present	.39	36.4	572	19.5
Range 14				
Discourse continuity				
Diff Subj	.56	54.0	1259	43.0
Same Subj & Diff TAM	.48	45.5	761	26.0
Same Subj & Same TAM	.43	40.9	910	31.1
Range 13				
Clause type				
Subordinate	.60	58.8	400	13.7
Main	.48	46.0	2530	86.3
Range 12				
TAM				
Preterit	.53	51.8	602	21.8
Imperfect	.53	51.3	398	14.4
Present	.48	54.2	1758	63.7
Range 05				
Verb class				
Perception	[.59]	56.5	85	02.9
Possession	[.56]	54.9	328	11.2
Cognition	[.51]	49.1	228	07.8
Other	[.50]	48.2	1375	46.9
Speech	[.48]	43.5	493	16.8
Relational	[.45]	43.0	421	14.4
Range *n.s.*				
Polarity				
Affirmative	[.51]	47.9	2632	89.8
Negative	[.45]	46.3	298	10.2
Range *n.s.*				
Morphological irregularity				
Regular	[.50]	48.0	1260	43.0
Irregular	[.50]	47.5	1670	57.0
Range *n.s.*				

Total Chi-square = 563.3672; Chi-square/cell = 1.3225; Log likelihood = -1967.935

5.4.1 Modal

The strongest factor group in affecting the realization of pronominal 3sg subjects in these data is modal. It can be seen in Table 31, the lack of a modal verb favors a pronominal 3sg subject (.53), whereas the presence of a modal favors an unexpressed 3sg subject (.39). This is the opposite to the direction of effect that we observed for 1sg and 2sg.

Five modals constitute 69% of all modals occurring with 3sg subjects, namely *ir* 'to go' (N=83, 52% pronominal), *poder* 'can' (N=47, 49% pronominal), *estar* 'to be' (N=44, 41% pronominal), *ter* 'to have' (N=37, 40% pronominal), and *querer* 'to want' (N=184, 10% pronominal). Four of these modals demonstrate a roughly even distribution between pronominal and unexpressed subjects. However, the modal *querer* 'to want' displays a much more skewed preference for unexpressed pronouns in these data, and note that it alone represents 32% (184/572) of the occurrences of modal verbs with 3sg subjects. And I would like to argue that it is because of *querer* 'to want' that we see such a strong effect of modals in 3sg subjects.

Another source of explanation for the patterning of modals in these data comes from its interaction with **TAM**. It can be seen in Table 32 below that the strong effect of 3sg unexpressed subjects and the realization of a modal is mostly observed in the **PRESENT** tense, with 29% realized pronominally and 49% without, while the imperfect showed rates of 51% with pronominal subjects and 51% without. This is the only place where we see that the data is not behaving evenly, thus suggesting that it has something to do with the pattern that emerged in the VRA.

Table 32. realization of 3sg subjects by TAM and modal.

		Present	Imperfect	Preterit
		P	P	P
Presence				
	N	105	32	31
	%	29	51	46
Absence				
	N	690	172	281
	%	49	51	53
Total				
	N	795	204	312
	%	45	51	52
χ^2		47.749*	.088**	.933**

* Significant at p<.005
** Not significant

Moreover, the modal *querer* 'to want', the most frequently occurring modal in these data, occurs 93% (N=171/184) of the time in the **PRESENT** tense. So, I will argue later in this work that the modal *querer* 'to want' forms a construction with 3sg unexpressed subjects and the present tense, and that this construction is clearly affecting the evolution of pronominal subjects with 3sg subjects.

5.4.2 Discourse Continuity

This factor group was the second strongest in conditioning the realization of 3sg pronominal subjects in these data. As can be seen in Table 31, as expected, and as we saw for 1sg (though not for 2sg) contexts of **LOW CONTINUITY** favor the realization of pronominal subjects (.56) while contexts of **HIGH CONTINUITY** disfavor the occurrence of pronominal subjects incrementally, (.48) and (.43).

As opposed to 1sg subjects that demonstrated specific patterns of **DISCOURSE CONTINUITY** and **TAM**, the data for 3sg subjects behaves very similarly in all three TAMs, which suggests that the effect is truly an artifact of discourse and not a result of an interaction with other factor groups. This can be observed in the data documented in Table 33 where in

all TAMs we can see a steady increase in rates of pronominal expression from **HIGH CONTINUITY** contexts to less continuous ones.

Table 33. Distribution of 3sg subjects by TAM and discourse continuity.

		Present	Imperfect	Preterit
		P	P	P
Same Subj & Same TAM				
	N	245	39	74
	%	41.0	37.9	42.5
Same Subj & Diff TAM				
	N	168	62	79
	%	42.6	48.8	48.2
Diff Subj				
	N	382	103	159
	%	50.0	61.2	59.3
Total				
	N	795	204	312
	%	45.2	51.3	51.8
χ^2		13.285*	14.500*	15.435*

* Significant at p<.005

5.4.3 Clause type

This factor group was the third strongest in conditioning the realization of 3sg pronominal subjects. As per Table 31, it can be seen that **SUBORDINATE** clauses (.60) strongly favor the realization of 3sg pronominal subjects as opposed to **MAIN** clauses, which disfavor the occurrence of pronominal 3sg subjects (.48). This is a pattern that has held true throughout all the analyses conducted so far, thus reinforcing the strong effect of this factor group on the variable as a whole.

5.4.4 TAM

This factor group is the weakest in the conditioning of 3sg expressed subjects in these data as evidenced by the range (05). The **PRETERIT** and the **IMPERFECT** slightly favor the use of pronominal subjects (.53) and the **PRESENT** disfavors them (.48). Moreover, if we look at the rates of expression for each **TAM** in Table 31 under the column '% of expressed' we can see

that all TAMs show an almost even distribution between pronominal and unexpressed subjects. This would likely account for such a weak effect of this factor group.

The findings noted for **TAM** with 3sg subjects suggest that there may be an effect of morphological ambiguity, except for the preterit. I would like to argue that there is not, rather what we observe is what Silva-Corvalán has argued to be the motivation of TAMs in discourse in that "es el tiempo verbal, más que la ambigüedad, y aun más específicamente la function de los diferentes tiempos en el discurso, lo que se correlaciona com expresíon del sujeto" (2001, p. 161).

The VRA shows that the past tenses favor expressed subjects (.53), while the **PRESENT** disfavors it (.48). When we look at the way these TAMs behave in discourse, it is observed that the **IMPERFECT** is used to indicate a pause in the sequence of events to describe some aspect that the speaker finds important, a well-recognized cross-linguistic pattern (Bybee, et al., 1994; Bernard Comrie, 1976; Hopper, 1979). Silva-Corvalán defends this interpretation as she states:

> Dada la funcíon pragmática del tiempo en el discurso, se podría esperar un número menor de sujetos expresos con Pretérito, y percentajes que irían en aumento con los verbos de tipo B [present] y C [imperfect]. Esta prediccíon se basa en el presupuesto de que los pronombres sujeto expresados atraen la atencíon hacia el referente del sujeto y consecuentemente disminuyen la atencíon prestada al evento (Silva-Corvalán, 2001, p. 162).

This distinction between these TAMs in regard to their function in discourse is a more reliable explanation for how these TAMs are patterning with pronominal expression. In adopting this explanation for how these TAMs are patterning with pronominal subjects, we are refuting a long tradition of resorting to formal explanations in favor of more functional explanations for these phenomena.

5.4.5 Summary

In this section the results for the VRA for 3sg subjects were presented. In this analysis only four of the seven predictors were selected as significant in the realization of 3sg pronominal subjects. These factor groups were **MODAL**, **DISCOURSE CONTINUITY**, **CLAUSE TYPE**, and **TAM**, respectively in order of magnitude of effect.

MODAL was the strongest factor group selected as significant in predicting the realization of 3sg pronominal subjects. The analysis shows that clauses without a modal were more likely to have an expressed 3sg subject, whereas clauses with modals were less likely to simultaneously contain such pronouns. The sources of explanation were offered to understand the behavior of this factor group with 3sg subjects. Firstly, it is argued that the high frequency with unexpressed subjects of one of the modals examined in the data strongly supports the non-realization of pronominal 3sg subjects. The modal *querer* 'to want' accounts for 32% of all modals that occur in these data, of which only 10% occur with pronominal subjects. Three other modals also show a slight favoring for occurring in clauses without a pronominal subject, namely *estar* 'to be', *ter* 'to have', and *poder* 'can'. These modals occur with clauses with pronominal subjects less than 50% of the time, with *ter* 'to have' occurring in these contexts only 40% of the time. When we add these three other verbs to *querer* 'to want', we observe a class of modals that occur frequently in clauses without pronominal subjects accounting for 55% of all modals in these data. This high frequency of these modals and their patterning with unexpressed subjects contribute to the pattern that emerges in these data.

Secondly, the patterning of these modals with 3sg subjects is complemented by their patterning with **TAM**, more specifically with the **PRESENT** tense. It shown that, in these data, the effect of modals appears to be limited to modals in the present tense, and in this case, to

querer ‘to want’ in the present tense. Thus, it is apparent that there is a construction effect at play here, and such effects will be discussed in more detail in Chapter 7.

The second factor group to contribute to a significant effect on the realization of 3sg pronominal subjects is **DISCOURSE CONTINUITY**. The findings observed for this factor group reinforce the hypothesis established earlier in this study and in the literature on subject expression, i.e., more continuous referents in clauses that have the same **TAM** as their previous clause are more likely to be realized as zero than less continuous referents that have a different **TAM**. This patterning is borne out in works on subject expression in Spanish and in Brazilian Portuguese, and they are also borne out here is this study. The patterning observed for 3sg subjects is very clearly discourse-based because when we attempt to cross tabulate this factor group with others, we observe the same pattern presented in the analysis. Thus, I would safely argue that the effect of **DISCOURSE CONTINUITY** in these data is independent from effects of other factor groups.

Turning now to the results concerning **CLAUSE TYPE**, the patterning observed in this dataset is consistent with those observed for 1sg and 2sg subjects, that is, **SUBORDINATE** clauses favor pronominal expression and **MAIN** clauses disfavor the occurrence of these subjects. I will discuss more fully the effect of this factor group in the data shortly (cf. section 5.5.1).

The last factor group to affect the realization of pronominal 3sg subjects in these data is **TAM**. A similar pattern was observed for the tenses in that both the **IMPERFECT** and the **PRETERIT** favor the occurrence of pronominal subjects. Inversely, the **PRESENT** tense tends to disfavor the occurrence of pronominal subjects. For this particular data I invoke the frequency of these TAMs in discourse as a possible explanation as to why they are behaving

in the way they are with pronominal subjects. To begin with, the **PRESENT** tense is the most frequent tense of the three in these data, which alone accounts for two thirds of the data analyzed. In addition, while the **PRESENT** shows a slightly higher percentage of occurrence of pronominal subjects than the other two tenses, the sheer number of unexpressed tokens that occur with the **PRESENT** overwhelm the total number of tokens that occur with the other two tenses, thus the effect of the **PRESENT** being so strong in these data. Moreover, taken that the pattern of pronominal expression has been considered to be more innovative in the language, it is also expected that it is going to have a difficult time penetrating the realm of present reference because of the frequency of the **PRESENT** in discourse. The same cannot be said for the past tenses, the **IMPERFECT** and the **PRETERIT**. These tenses are considerably less frequent than the **PRESENT**, accounting altogether for roughly one third of the data. Consequently, these tenses are more open to being affected by syntactic patterns that are becoming more frequent in the language as pronominal expression is.

Furthermore, the pattern of behavior with respect to **TAM** appears to be linked to the pattern observed for **DISCOURSE CONTINUITY**. While the **PRESENT** does not seem to be affected by the continuity of reference, showing an even rate of pronominal subjects across the three levels, the past tenses seem to be more affected by subject continuity with maximal continuity disfavoring the realization of pronominal subjects, as opposed to minimal continuity that favors pronominal subjects with the past tenses. Thus, while the argument for the high frequency of the **PRESENT** tense stands, it is not to be seen isolated from other factors, such as **DISCOURSE CONTINUITY**.

Another avenue of explanation for the patterning observed with **PRESENT** tense in these data resides in the patterns of 3sg subjects and predicates in the **PRESENT** tense and the

way they are patterning, individually, with pronominal expression. This explanation will be explored more fully in Chapter 7.

5.5 Comparison between the three subjects

In the preceding chapter and sections in this chapter I have presented the results for the overall analysis of subject expression as well separate analyses for each person to examine the different factors that condition the realization of expressed subjects. Table 34 below presents the results of the independent VRAs for subject expression in the 1sg, 2sg, and 3sg subjects' datasets. It is by comparing these results that we are able to identify underlying similarities and differences in the linguistic conditioning of pronominal expression across the three datasets, in order to determine whether the contrasting rates observed are representative of grammatical differences, or whether they should be attributed to other factors.

These results provide three pieces of evidence to help us understand the conditioning of the variation in subject expression (Bailey, 2002; Poplack & Tagliamonte, 2001). Firstly, those groups that have a significant effect on variant choice are distinguished from those that do not. As Table 34 shows, the three datasets behave precisely in the same way only in regards to **CLAUSE TYPE**, while all the other factor groups are not selected as significant in all three datasets. Secondly, the hierarchy of constraints for each factor group is observed in each dataset. Note that in these data, not all factors are behaving in the same way across the three datasets. Each will be discussed in more detail shortly. Thirdly, the magnitude of effect of the factor groups is determined. The single most striking observation to emerge from the data comparison was not so much the degrees of similarity between the three datasets, but the fact that not all factor groups selected as significant in one analysis were selected in the other two. What is more, even those factor groups that showed significance through different

analyses, they did not always demonstrate a similar hierarchy of constraints. In short, this comparison of the three datasets will reveal that these three persons are patterning very differently with pronominal expression.

The first comparison that can be established between the three data sets concern their distribution in terms of realization. 1sg subjects occur pronominally in two thirds of the data, while 2sg and 3sg subjects show a more balanced distribution; 2sg is expressed 53% of the time and 3sg 48% of the time.

Table 34. Comparison of Multivariate Analyses of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 1sg, 2sg, and 3sg subjects.

	1sg	**2sg**	**3sg**
Total *N*	3447	1689	2930
% *expressed*	64.7	53.4	47.7
Corrected *Mean*	.671	.547	.475
	Prob.	**Prob.**	**Prob.**
Verb class			
Possession	.62	.78	[.56]
Speech	.73	.48	[.48]
Other	.51	.70	[.50]
Relational	.48	.70	[.45]
Perception	.46	.34	[.59]
Cognition	.34	.24	[.51]
Range	39(1)	54(1)	*n.s.*
Clause type			
Subordinate	.76	.79	.60
Main	.45	.45	.48
Range	31(2)	34(2)	12(3)
Discourse continuity			
Diff Subj	.54	[.47]	.56
Same Subj & Diff TAM	.50	[.50]	.48
Same Subj & Same TAM	.41	[.52]	.43
Range	13(3)	*n.s.*	13(2)
TAM			
Imperfect	[.54]	.61	.53
Preterit	[.49]	.48	.53
Present	[.50]	.56	.48
Range	*n.s.*	13(3)	05(4)
Polarity			
Affirmative	.51	[.50]	[.51]

Negative		.45	[.47]	[.45]
	Range	0(5)	*n.s.*	*n.s.*
Modal				
Absent		[.50]	.48	.53
Present		[.53]	.58	.39
	Range	*n.s.*	10(5)	14(1)
Morphological irregularity				
Regular		.54	.45	[.50]
Irregular		.47	.56	[.50]
	Range	07(4)	11(4)	*n.s.*

In Table 34 we can see that **CLAUSE TYPE** is the only factor group to fare significantly, and with the same direction of effect, throughout the three different analyses. The remaining factor groups are selected as significant either with one of the subjects or with two of them at most. Each of the seven predictors included in these analyses is selected as significant for at least one person. In the following subsections I will discuss the patterning observed for these subjects for each of the seven predictors included in the analyses. To begin, I will now turn to the discussion of clause type.

5.5.1 Clause type

This factor group was selected as significant in all three analyses of the three data sets. It is the only factor group to play a conditioning role in the realization of pronominal subjects in all these data. It was selected as the second strongest factor group for 1sg and 2sug subjects and as the third strongest for 3sg subjects, and in all three analyses the direction of effect is the same: **SUBORDINATE** clauses strongly favor expressed subjects and **MAIN** clauses slightly disfavor them.

As was discussed earlier (section 4.1.2), this is a very surprising finding, and in fact, contrary to any hypotheses that may have been laid out. It has been widely noted that major

grammatical changes are normally transmitted from **MAIN** clauses to **SUBORDINATE** clauses and not vice versa (Givón 1976, p. 170-1).

A traditional explanation can be found in **DISCOURSE CONTINUITY**. When the subject and the **TAM** of the matrix clause are different from the subject and **TAM** of the **SUBORDINATE** clause, the latter tend to be mostly realized with pronominal subjects.

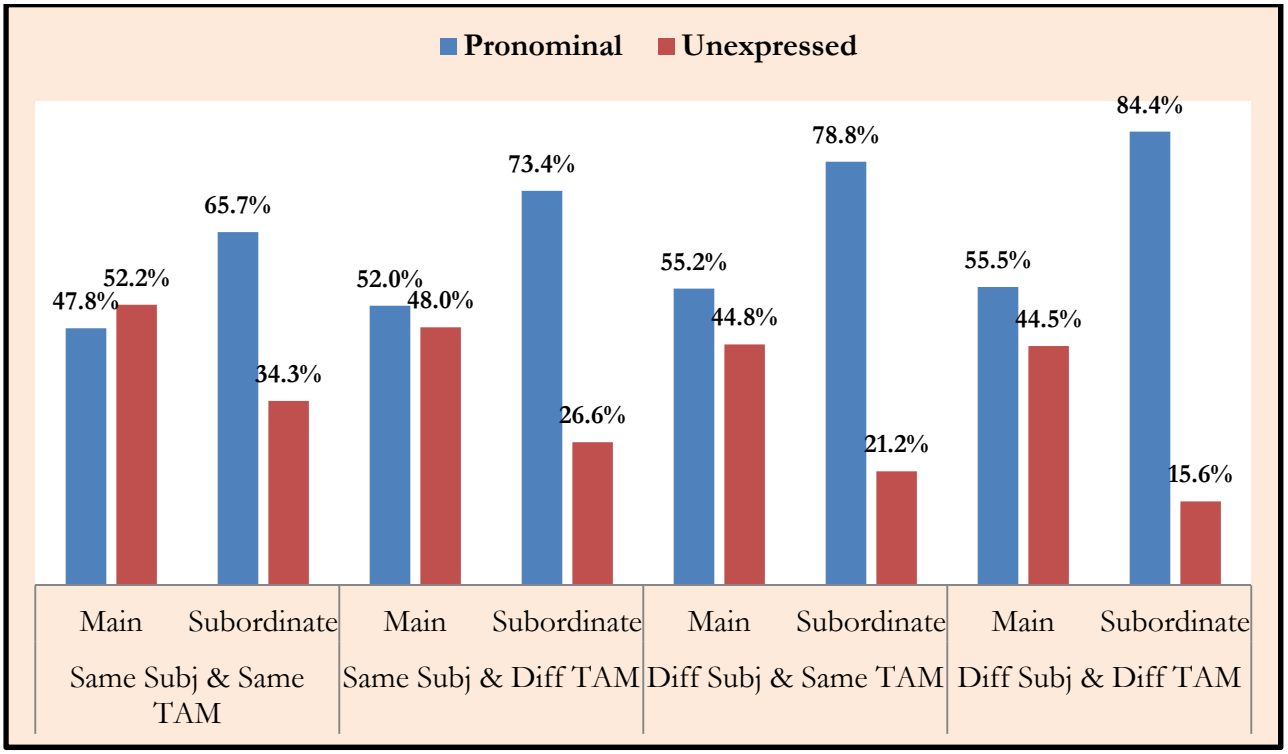

Figure 9. Distribution of pronominal expression by clause type and discourse continuity.
Same Subj. & Same TAM: Chi-Square = 28.494; p<.005
Same Subj & Diff TAM: Chi-Square = 44.900; p<.005
Diff Subj & Same TAM: Chi-Square = 47.632; p<.005
Diff Subj & Diff TAM: Chi-Square = 116.426; p<.005

Finally, while these results go against typological tendencies, it is not completely unheard of, nor it is the first of its kind. Detges reports that Old French from the 11th century onwards showed increased frequency of pronominal subjects in **SUBORDINATE** clauses (p. 77). It is known that nowadays French is an obligatory subject language, so the finding

observed here is indeed a possible path toward obligatory expression based on what has been found for French.

5.5.2 *Verb class*

Only 1sg and 2sg subjects appear to be conditioned by **VERB CLASS**. It is the strongest factor group in predicting the variable for these two persons. While this factor group does not have a significant effect on 3sg subject expression, we can nevertheless conduct a comparison based on the hierarchies of constraints of the three persons.

Table 35 provides the probability observed for each factor presented in the first column, the direction of effect in the second column, and the place in the hierarchy in the third column.

Table 35. Direction of effect for verb class across the three persons.

	1sg			2sg			3sg		
	Prob.	Direction	Hierarchy	Prob.	Direction	Hierarchy	Prob.	Direction	Hierarchy
Possession	.62	→	2	.78	→	1	[.56]	→	2
Speech	.73	→	1	.48		4	[.48]		5
Relational	.48		4	.70	→	3	[.45]	←	6
Perception	.46		5	.34	←	5	[.59]	→	1
Cognition	.34	←	6	.24	←	6	[.51]		3
Other	.51		3	.70	→	2	[.50]		4

→ Favors pronominal subjects
← Disfavors expression

From the table above we can see that only one factor are consistent across the three different analyses, namely **POSSESSION**. This factor reported a preference for co-occurring with pronominal subjects throughout the three analyses. **POSSESSION**, is indeed the strongest factor to condition expression in 2sg subjects and the second strongest in 1sg and 3sg subjects. This can be observed by examining the verbs in this class and their distribution with pronominal subjects illustrated in Table 36.

Table 36. Most frequent **POSSESSION** verbs for each person.

	1sg	2sg	3sg
	P	**P**	**P**
ceder 'to give'	--	--	0.00%
conseguir 'to get'	--	4 2.68%	4 2.22%
dar 'to give'	60 23.72%	21 14.09%	50 27.78%
entregar 'to deliver'	--	1 0.67%	--
receber 'to receive'	8 3.16%	1 0.67%	10 5.56%
ter 'to have'	185 73.12%	114 76.51%	107 59.44%
tirar 'to get'	--	5 3.36%	3 1.67%
trocar 'to exchange'	--	--	0.00%
vender 'to sell'	--	3 2.01%	6 3.33%
Total	**253**	**149**	**180**

1sg chi-square = 9.157, p<.05
2sg chi-square = 6.194, p<.05
3sg chi-square = 3.528, p=.171

What is more, the effect of this category as a whole can be attributed to one verb in particular, namely *ter* 'to have'. This verb alone accounts for 65% of all **POSSESSION** predicates (248 tokens with 1sg with 77% pronominal, 131 tokens with 2sg with 87% pronominal, and 181 tokens with 3sg with 60% pronominal), of which 73% occur with pronominal subjects. Thus, if the most frequent verb in this category has such high rates of expression, it is expected that the rest of the category, which are of much lower frequency and thus more susceptible to being affected by other more frequent patterns, may follow the same pattern and occur more promptly with pronominal subjects.

The category of **'OTHER'** that consists of verbs that did not fit any of the categories established in this work also favors expression in all three persons. This is most interesting because it suggests that most verbs in the data are probabilistically more likely to be realized with pronominal subjects. These findings point to the possibility that pronominal expression is the preferred pattern in the language except for the verbs that fall into the categories of **POSSESSION**, **SPEECH**, **RELATIONAL**, **PERCEPTION**, and **COGNITION**. Given that the verbs in these categories are high in token frequency, they play an important role in the way the pattern is preserved in the language. The category **'OTHER'**, on the other hand, consists of much higher type frequency than the other categories combined, which induces us to assume that the remaining verbs of the language are complying with the pattern of pronominal expression more promptly.

Returning to the results illustrated in Table 36, we observe that **SPEECH** highly favors expression in 1sg subjects but has a weak effect in 2sg and 3sg subjects. This favoring of **SPEECH** predicates for pronominal subjects with 1sg is a reflection of the patterning observed with the verb *dizer* 'to say' and *falar* 'to speak' discussed earlier. Morales (1997), Enriquez (1984), and Travis (2007) also found this category of verbs to favor pronominal expression in Spanish as well with high rates of pronominal expression.

The remaining three categories show very distinct patterns throughout the three persons. **RELATIONAL** predicates highly favor pronominal subjects with 2sg, is the factor that most disfavors expression with 3sg subjects, and has little effect with 1sg subjects. **PERCEPTION** and **COGNITION** predicates show similar patterning with 1sg and 2sg subjects, with a much stronger effect for both with 2sg subjects. With these subjects, these two verb classes highly disfavor expressed subjects. **PERCEPTION** predicates favor pronominal

subjects with 3sg subjects. These findings for **PERCEPTION** and **COGNITION** verbs are not at all surprising for the three persons. It has been found and claimed again and again that verbs of **PERCEPTION** and **COGNITION** are highly frequent when occurring with 1sg and 2sg subjects and much less frequent when occurring with 3sg subjects. If we accept the fact that BP is moving toward obligatory pronominal expression, then it may be posed that **COGNITION** verbs from the very beginning of the language were occurring mostly, if not nearly categorically, with non-expressed subjects.

Now, as was discussed in 5.2.1 and 5.3.1, **COGNITION** verbs followed by **PERCEPTION** ones are the most frequently occurring verbs with 1sg and 2sg subjects. Thus, it is expected that these predicates would be most resistant to the introduction of a new syntactic pattern between these verbs and their 1sg and 2sg subjects. In short, pronominal expression may be frequent and more favored in other contexts, but it still has long ways to go in order to overcome the constraints imposed by this verb class.

A similar argument with a different direction of effect can be observed for **PERCEPTION** and **COGNITION** verbs with 3sg subjects. These verbs form a low-frequency niche within the group of verbs that occur with 3sg subjects. Thus, **PERCEPTION** and **COGNITION** verbs are more susceptible to being influenced by patterns such as pronominal expression.

It can be concluded then that **PERCEPTION** and **COGNITION** predicates form the loci of constraint of pronominal expression for 1sg and 2sg subjects and the loci for the frequency of pronominal expression to increase within 3sg subjects.

5.5.3 *Morphological Irregularity*

This factor group was selected as significant in the analyses for 1sg and 2sg subjects. while 3sg subjects do not show a significant effect of this factor group, it is also to be noted that the factors did not show a direction of effect.

For 1sg subjects we see that **REGULAR** verbs favor pronominal subjects while **IRREGULAR** verbs disfavor them. 2sg subject show the reverse effect whereby **REGULAR** verbs disfavor pronominal subjects and **IRREGULAR** verbs favor pronominal expression. Such disparity between the two datasets reinforce the premise of this work, which is to show that these subjects need to analyzed separately. Secondly, there seem to be underlying lexical effects with each of these persons that may be guiding the patterns observed in the statistical analyses. This hypothesis will be investigated in more detailed in Chapter 6.

5.5.4 *Discourse Continuity*

This factor group only emerges as significant in the analyses for 1sg and 3sg subjects. This is somewhat expected due to the nature of the three persons in discourse. While 1sg and 3sg subjects can demonstrate high levels of continuity in discourse, the same cannot be expected from 2sg subjects which do not tend to occur in contexts of **HIGH CONTINUITY** as often as the other two persons. In fact, contexts of **HIGH CONTINUITY** only accounts for one third of the data for 2sg subjects versus nearly half the data for the other two subjects.

The direction of effect of this factor group is essentially the same for both 1sg and 3sg subjects, the greater the discourse continuity (measured both in terms of continuity of subject and TAM), the more likely they are to be realized without pronominal subjects. On the other hand, the lesser the discourse continuity, the more likely they are to be realized pronominally.

Indeed, in both persons, but more strongly in 1sg subjects, the effect is truly only observable within past TAMs as documented in Table 37. This table presents two statistical tests to help us understand the distribution of the data at hand. The first test to be illustrated in this table is the chi-square test, which illustrates the level of significance; in other words, it tells whether or not these values are due to chance or to a correlation between the variables being tested. There are two levels of significance illustrated in this table, the first is a p value below .05 and the second, and stronger, p value below .005. However, chi-square tests do not give us the direction of effect, but it only gives significance of the distribution, thus we are using the R value to reveal the direction of effect. The second test is the Pearson's R, which represents the direction of effect. A positive value favors pronominal subjects and a negative value favors non-expression. The higher the value in comparison to one another, the stronger the effect.

Table 37. Crosstabulation of discourse continuity and TAM across all persons.

		Present			**Imperfect**			**Preterit**		
		1sg	**2sg**	**3sg**	**1sg**	**2sg**	**3sg**	**1sg**	**2sg**	**3sg**
Same Subj & Same TAM										
Pronominal	N	313	157	245	45	12	39	95	12	74
	%	59.3	61.8	41.0	54.2	63.2	37.9	46.3	52.2	42.5
Same Subj & Diff TAM										
Pronominal	N	172	67	168	112	18	62	130	23	79
	%	62.8	49.3	42.6	74.7	78.3	48.8	66.7	62.2	48.2
Diff Subj & Same TAM										
Pronominal	N	331	217	186	23	14	31	89	26	52
	%	61.1	57.4	52.0	51.1	58.3	60.8	71.2	59.1	55.9
Diff Subj & Diff TAM										
Pronominal	N	342	156	196	175	36	72	262	57	107
	%	63.3	34.5	48.0	83.7	75.0	61.5	75.7	48.3	62.6
Pearson's R*		.027	-.189	.071	.201	.053	.185	.226	-.074	.160
Chi-square**		2.094	65.337*	13.285*	38.125*	3.307	14.500*	51.515*	2.952	15.435*

* Significant at p<.005.

For example, for 1sg subjects, the **PRETERIT** has a higher R then the **IMPERFECT**, and both values have a positive value suggesting that the preterit has a stronger effect on the conditioning of pronominal 1sg subjects. This test also shows the degree of correlation between pronominal expression, **TAM**, and **DISCOURSE CONTINUITY**. Using the same example of the comparison between the **PRETERIT** and the **IMPERFECT** with 1sg subjects, we see that the increase in pronominal expression as continuity decreases is much stronger in the **PRETERIT** than it is in the **IMPERFECT**. This particular finding indicates that the notion of ambiguity within **TAM** alone cannot fully explain the pattern of pronominal expression with TAMs. These tests should provide us with the evidence for the strong interaction between **DISCOURSE CONTINUITY** and **TAM**.

The most striking result to emerge from this table is that both 1sg and 3sg subjects are patterning very similarly in regards to the patterning of pronominal expression with discourse continuity. In these two subjects, we see a clear increase in pronominal realization with a decrease in continuity. What is more, this pattern is almost neutralized in the **PRESENT**, suggesting that the effects of this factor group can only be fully analyzed in the past TAMs.

5.5.5 Polarity

This factor group was selected as significant in the analyses for 1sg and 3sg, with the same direction of effect for all three persons, subjects in these data. However, this factor group was a weak predictor of the realization of pronominal subjects based on the range in comparison to the other factors selected as significant in each analysis for these two persons.

While this factor group was acknowledged as significant by the analysis presented in Duarte (1993), she does not offer an explanation for the reason why this should be the case. In an earlier section of this work (5.2.5), I offered the claim that there must a close

relationship between the negation markers and the verbs that induce the non-occurrence of pronominal subjects in these two datasets. This is actually borne out when we look at the most frequent construction to occur in these two datasets. Table 38 provides the breakdown of the most frequently occurring combinations of negation + predicate in the two datasets for 1sg and 3sg subjects.

Table 38. Distribution of construction with *saber* 'to know' across 1sg and 3sg subjects.

		1sg	3sg
		P	P
não sabe '(you) don't know'			
	N	37	-
	%	36.5	
não tem '(it) doesn't have'			
	N	-	11
	%		22
num tenho '(I) don't have'			
	N	23	-
	%	66.5	
não é '(it) isn't'			
	N	-	9
	%		10
num sou '(I) am not'			
	N	11	-
	%	69	
num dá '(it) doesn give'			
	N	-	1
	%		09
	Total	71	21
		42	14

It must be first noted that these few constructions account for a large portion of negative statements in the two datasets. These constructions correspond to 34% of the data of 1sg subjects and 52% of the data of 3sg subjects. In both sets of constructions we also observe that they favor non-expression, with a stronger favoring in the dataset of 3sg subjects. Such difference in strength cannot be captured in the VRA, thus the importance of breaking it down. But what is really striking in the two datasets is that all constructions

disfavor pronominal expression in the 3sg subjects' dataset, while only two constructions[36] in the 1sg subjects' dataset disfavor pronominal subjects. The remaining constructions slightly favor the occurrence of pronominal subjects. Thus, it can be suggested that the patterning of negation without pronominal subjects is likely to be an artifact of these two constructions with the verb *saber*, which disfavor pronominal subjects as a whole. It is the high frequency of these constructions without pronominal subjects that render negative statements more likely to occur without pronominal subjects.

5.5.6 TAM

The results for this factor group were observed to have a significant effect in the datasets for 2sg and 3sg subjects. Strong evidence of the favoring of pronominal subjects to occur with the **IMPERFECT** tense was found in both datasets[37], with a stronger effect in comparison to the other two tenses observed in the 2sg subjects' dataset. The **PRETERIT** and the **PRESENT**, contrariwise, showed opposing effects in the two analyses in that the former favored pronominal subjects with 3sg subjects but disfavored them with 2sg subjects (as with 1sg subjects).

While the patternings observed for the **PRETERIT** and the **PRESENT** may raise questions regarding the effect of this factor group on the variant, its patterning with the **IMPERFECT**, on the other hand, offers the most interesting finding in this factor group. This is the case because the **IMPERFECT** is the locus of **TAM** ambiguity in this data, thus its favoring

[36] These are going to be treated as two constructions to respect the ongoing grammaticization of *não* into *num* in BP. While the two forms appear to be different, their difference is only orthographic. The form *num* is a phonetically eroded form of *não* that is increasing in frequency in BP at present. So, even though the two constructions are behaving similarly in these data, I will consider them as two separate forms.

[37] The direction of effect is the same for the three persons.

of co-occurrence with pronominal subjects appears to support this claim. However, I will invoke the same explanation offered for this effect with 3sg subjects, in that this is more of frequency effect than an effect of the morphology of these TAMs.

It has been found again and again that lower token frequency forms tend to be affected more quickly by emerging patterns of increased frequency than high frequency forms. This can be observed by examining the type-token ratio (TTR) for each **TAM** according to expression. Table 39 illustrates these ratios. The closer the TTR is to 0, the less diversity of types it shows in the data, the inverse being true, the closer the TTR is to 100, the more diverse of types are occurring with the form. Thus, a TTR closer to 100 demonstrates the effects of a more productive pattern.

Table 39. TTR ratios of pronominal subjects for all persons across different TAMs

	Expressed				**Unexpressed**			
	N	*%*	*types*	*TTR*	*N*	*%*	*types*	*TTR*
1sg								
Present	1245	66.1	125	10	639	33.9	83	13
Preterit	536	61.5	112	20.9	335	38.5	78	23.3
Imperfect	355	72.9	81	22.8	132	27.1	34	25.8
2sg								
Present	649	53.2	136	21	571	46.8	78	13.7
Preterit	109	48.9	31	28.4	113	51.1	7	06.2
Imperfect	80	70.2	26	32.5	34	29.8	6	17.6
3sg								
Present	795	45.2	200	25.2	963	54.8	272	28.2
Preterit	312	51.8	72	23.1	290	48.2	105	36.2
Imperfect	204	51.3	51	25	194	48.7	85	43.8
Total								
Present	2777	57	328	11.8	2085	43	339	16.3
Preterit	956	56	153	16.00	739	44	157	21.24
Imperfect	639	64	130	20.34	360	36	109	30.28

What the table above is really showing is that although pronominal expression is the favored pattern in some of these TAMs, it is not the most diverse pattern in terms of types with which it occurs. To take the dataset of 3sg subjects as an example, we see that in terms

of percentages the imperfect shows similar rates for pronominal expression (51.3%) and unexpressed subjects (48.7%), but this favored patterned is not the most productive in these data. When we look at the TTR, we still see that non-expression is much more productive in the data with about 1.5 times as many verb types occurring without a pronominal subject than with one (43 vs. 25). This higher TTR for unexpressed subjects is observed throughout all TAMs in the dataset for 1sg and for 3sg subjects. 2sg subjects, on the other hand, demonstrate the opposed pattern in that all tenses show a higher TTR with pronominal subjects suggesting that this pattern is more productive in these data. Although 2sg subjects demonstrate a preference for pronominal subjects in the TTR, the overall patterning of the language as evidenced in the TTRs for the total is that non-expression is still a more productive pattern in the language even if the percentages may point to an increase in rates of subject expression.

5.5.7 Modal

In the presence of modals, we again see distinct patterning. The presence of modals disfavors 3sg subject expression, but favors 1 and 2sg expression. In addition, this factor group has differing levels of magnitude across the two analyses as well. Modal is the strongest factor group in conditioning pronominal expression with 3sg subjects, while it is one of the weakest predictors of pronominal expression with 2sg subjects, and is not significant with 1sg.. Such variability reinforces the claim that different persons must be examined separately if we as linguists are to attain a better understanding of their patterning.

But how can we explain these conflicting directions of effect? If we look at the way modals are patterning with predicates with each of these subjects, we can see that local

patterns of subject, modal, and predicate are responsible for such conflicting results and that a generalized statistical model cannot capture these intricacies.

For example, the dataset for 3sg subjects documents a total of five modals which account for 69% of all modals occurring with 3sg subjects. 3sg subjects show an overall rate of expression of 48%, and four of these modals are close to this average: namely *ir* 'to go' (43/83, 52% pronominal), *poder* 'can' (24/47, 49% pronominal), *estar* 'to be' (41% pronominal, 18/44), and *ter* 'to have' (15/37, 40% pronominal), while one modal in particular shows an almost categorical preference for unexpressed subjects, i.e. *querer* 'to want' (10% promominal, 20/184). This modal alone represents 32% (184/591) of all modals with 3sg subjects in these data. Thus, it is not surprising that its presence inhibits pronominal expression. This is a clear example of how a lexical form can impact a syntactic pattern such as pronominal expression. The other members of this class of modals occurring with 3sg subjects, while not as frequent as *querer* 'to want', are influenced by it because of its structural proximity and are thus more likely not to take a pronominal subject with 3sg subjects.

The inverse pattern can be observed with 2sg subjects. The pattern observed in these data is a favoring of pronominal expression by modals. As opposed to the data for 3sg subjects where we see an overwhelming effect of one modal in particular, in this dataset we see three modals behaving nearly identical in their pattern of occurrence with pronominal subjects, namely *ir* 'to go' (89% expression, 75/84), *poder* 'can' (N=41/45), and *ter* 'to have' (N=21/25). These forms account for 49% of all modals in these data (N=154/311), of which 89% (137/154) occur with pronominal subjects. Indeed, we can see that this dataset also falls

prey to the effects of individual lexical items that are behaving in a particular manner, thus contributing to general results that do not reveal the intricacies of the pattern.

5.6 Discussion and Summary

The analyses presented in this chapter confirm that to truly understand the nature of the variation between pronominal and unexpressed subjects in Brazilian Portuguese, one must conduct in-depth analyses considering local patterns that may emerge between subjects, verbs, TAMs, and other aspects in the clause. Moreover, they also invoke the need to identify and examine the several lexical, local effects that may be gearing the data toward a different direction than that of the part of the dataset that is behaving more according to constraints imposed on the syntactic pattern.

The discussion of the results for all factor groups included in these analyses assists in our understanding of their role in the conditioning of pronominal expression. Firstly, it is observed that pronominal expression is not conditioned by the same predictors across the three subjects. This is an innovative contribution of this study in that, in this chapter, we had the opportunity to examine each person separately along with a comparative look at their analyses. Indeed, the differences in the behavior of the factor groups can be considered steep in that an overall analysis with all the three persons combined loses its merit for it overshadows such peculiarities that emerged in the separate ones.

Secondly, not all factor groups fare significantly at the same magnitude of effect across different persons. For example, the analysis for 2sg subjects shows a much more pronounced difference between factor groups in terms of their strength of effect than the analysis for 1sg subjects, and even more than the one observed for 3sg subjects. In fact, the analysis for the latter shows that although the predictors selected as significant have an effect

on the variant, the strength of the effect is not very strong. For example, the strongest significant factor group in this analysis is **MODAL**, which shows a Range of only 14. This is the strength of effect of this group, that is, this is difference between the probability that most favors expression to the ones least favors it. However, in 2sg subjects, **VERB CLASS** is the strongest one and it shows a range of 54. In other words a much stronger effect in the conditioning of the variant or much more reliable predictor. In short, it seems that the factor groups adopted for these analyses are good predictors of pronominal expression for 1sg and 2sg subjects, but not as good predictors of pronominal expression for 3sg subjects.

6 FREQUENCY EFFECTS

6.1 Introduction

Frequency of use, especially relative frequency, is a crucial tool to aid the understanding of how grammatical units are organized in discourse. In this study, the frequency of types according to each of the persons was analyzed. Using this measurement allows us to examine the effect of those types that most occur with each of the subjects, rather than using a total token frequency which would obscure possible patternings.

Following this measurement of relative frequency, every predicate accounting for one percent or more of the data for the three persons was culled from each dataset. This is an arbitrary cutoff in terms of what can be considered as high frequency. However, there were two reasons why this cutoff point was chose. One is based on other scholars who seem to find such measure an adequate one (cf. Goddard, 2005 inter alia). In addition, preliminary analyses were conducted to test the effects of each different cutoff point, and there were no statistical differences on the results obtained by raising the cutoff points, thus I decided to extract these verbs that met the one percent cutoff point[38]. The verbs extracted from each dataset are presented in Table 40. It is interesting to observe that each person is patterning differently with the various predicates that occur in each dataset. Such patterning only

[38] Another set of statistical analyses that were performed and contributed to this decision of using the one percent as the cutoff point lies in the decision of whether to discriminate verbs that form constructions and those that do not. Tests on these high-frequency verbs that showed a greater percentage of construction use did not yield any statistical difference in the results presented here. Thus, it is my decision to show a more general, inclusive, result to test lexical as well as construction effects on pronominal expression.

reinforces our claim that these subjects must examined in terms of their patterning with predicates.

Table 40. Most frequent verbs occurring with each person.

1sg		2sg		3sg	
Verb	**%**	**Verb**	**%**	**Verb**	**%**
saber 'to know'	12	*saber* 'to know'	12.7	*dizer* 'to say'	12.4
achar 'to think'	9.3	*entender* 'to understand'	9.9	*ser* 'to be'	8.8
dizer 'to say'	8.4	*ter* 'to have'	7.7	*ter* 'to have'	6.2
ter 'to have'	7.2	*ver* 'to see'	7.6	*fazer* 'to do'	5.4
ser 'to be'	4.5	*olhar* 'to look'	7.3	*estar* 'to be'	4.5
fazer 'to do'	4.4	*fazer* 'to do'	3.8	*dar* 'to give'	3.4
falar 'to speak'	2.8	*ser* 'to be'	3.3	*querer* 'to want'	2.9
ver 'to see'	2.5	*achar* 'to think'	2.3	*saber* 'to know'	2.3
dar 'to give'	2.3	*estar* 'to be'	1.9	*falar* 'to speak'	1.9
estar 'to be'	2.2	*dar* 'to give'	1.8	*ir* 'to go'	1.8
ir 'to go'	2.1	*dizer* 'to say'	1.8	matricular	1.5
gostar 'to like'	1.7	*ir* 'to go'	1.5	*ficar* 'to stay'	1.4
ficar 'to stay'	1.4	*passar* 'to pass'	1.4	*sair* 'to leave'	1.4
querer 'to want'	1.4	*comprar* 'to buy'	1.4	*achar* 'to think'	1.2
tomar 'to take'	1.2	*pegar* 'to take'	1.2	*ver* 'to see'	1.1
pensar 'to think'	1.1	*ficar* 'to stay'	1.2	*passar* 'to pass'	1.0
lembrar 'to remember'	1.1	*pensar* 'to think'	1.2	Total:	57.2
trabalhar 'to work'	1.0	*deixar* 'to let'	1.1		
Total:	66.6	*falar* 'to speak'	1.1		
		imaginar 'to imagine'	1.1		
		querer 'to want'	1.1		
		Total:	72.4		
Other verbs	33.4	Other verbs	27.6	Other verbs	42.8

All these verbs were excluded from the VRAs for each person to test whether their absence would promote the model to better predict the realization of pronominal subjects. Each set of verbs will be discussed in their respective sections in Chapter 7. I will now turn to discussion of the analyses without these verbs.

6.2 1sg Subjects

After removing the high frequency verbs from this analysis, only 34% of the data was submitted to a new VRA to test the effect of the seven predictors analyzed in this study so far on variant choice. Results for the VRA without the most frequent verbs are presented in Table 41 and the comparison between this new analysis and the analysis with all verbs included is presented in Table 42.

The first point that needs to be made about this analysis concerns the change in rate of pronominal expression. This VRA shows an increase in pronominal expression of seven percent when compared to the analysis with all verbs included (65%, vs. 72% here) (see Table 16).

Another point that should be raised concerns morphological irregularity. The exclusion of the most frequent verbs in this dataset has apparently neutralized the effect of this factor group on variant choice, such that it is no longer significant. This is partly attributed to the fact that most of these verbs that were excluded from this analysis were indeed irregular verbs, which disfavored the occurrence of pronominal subjects in the data. Now that they are not part of the analysis, we see that the distinction between regular and irregular predicates is no longer applicable and the direction of effect is reversed.

Table 41. Multivariate Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 1sg subjects without the most frequent verbs.

Total *N*	1408			
% *expressed*	71.8			
Corrected *Mean*	.743			
	Probability	**% expressed**	**N**	**% data**
Clause type				
Subordinate	.77	89.7	234	16.6
Main	.44	68.2	1174	83.4
Range 33				
Verb class[39]				
Speech	.67	84.7	425	30.2
Other	.42	66.2	794	56.4
Cognition	.42	66.1	189	13.4
Range 25				
Discourse continuity				
Diff Subj	.54	76.5	759	53.9
Same Subj & Diff TAM	.48	69.9	302	21.4
Same Subj & Same TAM	.42	63.1	347	24.6
Range 12				
Morphological irregularity				
Irregular	[.54]	80.7	424	30.1
Regular	[.48]	68.0	984	69.9
Range *n.s.*				
TAM				
Imperfect	[.55]	76.6	201	15.6
Preterit	[.51]	72.3	650	50.3
Present	[.46]	68.2	440	34.1
Range *n.s.*				
Polarity				
Negative	[.53]	72.0	137	09.7
Affirmative	[.50]	70.1	1271	90.3
Range *n.s.*				
Modal				
Present	[.51]	73.8	290	20.6
Absent	[.50]	71.3	1118	79.4
Range *n.s.*				

Total Chi-square = 254.5975; Chi-square/cell = 1.1520; Log likelihood = -772.968

[39] Once the most frequently occurring verbs with 1sg subjects are removed from the data, the three factors laid out here are the only ones that remain.

A second point to be noted regarding this new analysis relates to the number of factor groups selected as significant, three as opposed to five in the original analysis. In this analysis only clause type, discourse continuity, and verb class appear to have an effect on the realization of the variable. And while there are no changes in the direction of effect within the significant factor groups, it must be noted that verb class only contains three classes now, speech, other, and cognition. And we observe that the classes speech and cognition maintain their patterns of occurrence with pronominal and without, respectively. The category other still favors pronominal expression, relatively, less than speech and more than cognition. In the analysis with all predicates included, this factor showed a slight favoring for the occurrence of pronominal subjects. In this analysis, however, it shows a favoring for the non-occurrence of these subjects. I attribute this change in direction of effect to the strong patterning that speech and cognition predicates have in the pattern of subject expression. By looking at the percentage of pronominal realization in the category of other, we see that this category indeed favors pronominal expression, but when compared to speech predicates, this favoring is much weaker.

In short, this analysis offers what I believe to be a more accurate portrayal of the behavior of pronominal expression in these data. When we exclude the most frequent verbs from the data, the pattern observed for pronominal expression with 1sg subjects can be summarized as follows:

- Pronominal subjects are favored with subordinate clauses, speech predicates, and in contexts of minimal discourse continuity (different subjects and different TAMs).

Table 42. Comparison of Multivariate Analyses of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 1sg subjects.

		1sg (all tokens)	1sg (high-frequency excluded)
Total *N*		3447	1408
% *expressed*		64.7	71.8
Corrected *Mean*		.671	.743
		Prob.	**Prob.**
Verb class			
Speech		.73	.67
Possession		.62	-
Other		.51	.42
Relational		.48	-
Perception		.46	-
Cognition		.34	.42
Range	39		25(2)
Clause type			
Subordinate		.76	.77
Main		.45	.44
Range	31		33(1)
Discourse continuity			
Diff Subj		.54	.54
Same Subj & Diff TAM		.50	.48
Same Subj & Same TAM		.41	.42
Range	13		12(3)
TAM			
Imperfect		[.54]	[.55]
Present		[.50]	[.51]
Preterit		[.49]	[.46]
Range	*n.s.*		*n.s.*
Polarity			
Affirmative		.51	[.50]
Negative		.45	[.53]
Range	06		*n.s.*
Modal			
Present		[.53]	[.50]
Absent		[.50]	[.51]
Range	*n.s.*		*n.s.*
Morphological irregularity			
Regular		.54	.48
Irregular		.47	.54
Range	07		06(4)

These results corroborate previous findings for Spanish and Portuguese with the exception that it accounts for patterns of subjects, predicates, and TAMs that emerge in discourse. These patterns are behaving differently from the rest of the language and they must be analyzed separately for their pattern combined does not necessarily amount to a pattern similar to the one observed in the analyses so far. In short, the analysis for 1sg subjects confirms the hypothesis that high frequency forms pattern differently than other forms in the language.

6.3 2sg Subjects

After removing the high frequency verbs from this analysis, only 27% of the data was submitted to a new VRA to test the effect of the seven predictors analyzed in this study so far on variant choice. Results for the VRA without the most frequent verbs for 2sg subjects are presented in Table 43 and the comparison with the analysis that includes all verb forms is presented in Table 44.

The results obtained for the analysis of 2sg subjects without the most frequent verbs occurring with these subjects reveal three important findings. Firstly, the rate of expression observed in the two different analyses increases by one quarter in the analysis presented in Table 43, from 53% expression with all verbs to 67% excluding the most frequent verbs. This is a greater proportional increase than that we saw for 1sg, which may be expected since a larger portion of the data have now been removed.

Table 43. Multivariate Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 2sg subjects without the most frequent verbs

Total *N*	463			
% *expressed*	67.2			
Corrected *Mean*	.691			
	Probability	**% expressed**	**N**	**% data**
Clause type				
Subordinate	.74	86.6	97	21.0
Main	.43	62.0	366	79.0
Range 31				
TAM				
Preterit	.65	77.4	62	15.6
Imperfect	.60	75.9	29	07.3
Present	.46	61.8	306	77.1
Range 19				
Discourse continuity				
Diff Subj	.53	72.6	292	73.4
Same Subj & Diff TAM	.42	57.5	106	26.6
Same Subj & Same TAM	--			
Range 11				
Verb class				
Other	[.52]	68.3	382	82.5
Cognition	[.49]	64.4	45	09.7
Speech	[.35]	58.3	36	07.8
Range *n.s.*				
Modal				
Absent	[.60]	76.7	343	74.1
Present	[.46]	63.8	120	25.9
Range *n.s.*				
Polarity				
Affirmative	[.50]	67.5	431	93.1
Negative	[.48]	62.5	32	06.9
Range *n.s.*				
Morphological irregularity				
Irregular	[.52]	67.6	56	12.1
Regular	[.50]	64.3	407	87.9
Range *n.s.*				

Total Chi-square = 108.3176; Chi-square/cell = 1.0415; Log likelihood = -269.163

The repercussion of this finding is twofold. Firstly, I will argue, as I did for 1sg subjects, that these high frequency verbs are forming their own individual patterns of variable

subject expression. On the flip side, the data without the most frequent verb forms show a much stronger favoring for the occurrence of pronominal subjects.

A second important finding revealed by this analysis is found in the results for TAM in that there is a change in the ordering of constraints. In the analysis with all verb forms included, the hierarchy of constraints presented was that the imperfect most favored pronominal expression, followed by the present, with the preterit the TAM that most disfavored expression. Without these high frequency verbs, while the imperfect and the present have maintained their relative rankings, the preterit is now the TAM that most favors expression. This is more in accordance with the patterns observed in other studies of subject expression in that past forms are more prone to occurring with pronominal subjects.

A third key difference observed in this analysis is not seen among the factors selected as significant, but among the ones that were not. Recall that in the analysis of 2sg subjects with all verbs included, verb class was the strongest factor group conditioning the realization of pronominal 2sg subjects. It was one and half time stronger than the next strongest factor group. In the analysis without the most frequent verbs, verb class does not fare as a significant factor in the conditioning of pronominal subjects, which leads us to believe that the lexical effects observed in the earlier analysis can be attributed, if not fully but at least in part, to the high frequency verbs excluded from the analysis.

Table 44. Comparison of Multivariate Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 2sg subjects.

		2sg – All tokens		2sg – High frequency verbs excluded
Total *N*		1689		463
% *expressed*		53.4		67.2
Corrected *Mean*		.547		.691
		Prob.		**Prob.**
Verb class				
Possession		.78		-
Relational		.70		-
Other		.70		[.52]
Speech		.48		[.45]
Perception		.34		-
Cognition		.24		[.49]
Range	54		*n.s.*	
Clause type				
Subordinate		.79		.74
Main		.45		.43
Range	34			31(1)
Discourse continuity				
Same Subj & Same TAM		[.52]		.42
Same Subj & Diff TAM		[.50]		-
Diff Subj		[.47]		.53
Range	*n.s.*			11(3)
TAM				
Imperfect		.61		.60
Present		.56		.46
Preterit		.48		.65
Range	13			19(2)
Polarity				
Affirmative		[.50]		[.50]
Negative		[.47]		[.48]
Range	*n.s.*		*n.s.*	
Modal				
Present		.58		[.50]
Absent		.48		[.46]
Range	10		*n.s.*	
Morphological irregularity				
Irregular		.56		[.52]
Regular		.45		[.50]
Range	11		*n.s.*	

The results obtained for the analysis of the conditioning of 2sg subjects without the most co-occurring frequent verbs offer valuable insight to the way these subjects are realized in discourse. These findings reiterate the claim established in the beginning of this work and in the previous discussion of 1sg subjects that these high frequency verbs form individual patterns that do not converge into a probabilistically analyzable more general one. Thus, the structure of pronominal expression in BP appears very dependent on the way these high frequency verbs behave.

A last note on the way these factor groups are behaving with 2sg subjects without the most frequent verb regards the pattern observed for **DISCOURSE CONTINUITY**. While this factor group was not selected as significant in either analyses, it is interesting to see that when the highest frequency verbs are excluded from the analysis this factor groups begins to behave the way it is predicted to behave – based on previous findings and the patterning observed with 1sg and 3sg subjects – in that same subjects and same TAMs tend to disfavor pronominal subjects while a change in subjects favor pronominal subjects. This is again strong evidence for this kind of analysis where lexical frequency is accounted for in how it affects more global syntactic patterns.

6.4 3sg Subjects

After removing the high frequency verbs from this analysis, only 44% of the data was submitted to a new VRA to test the effect of the seven predictors analyzed in this study so far on variant choice. Results for the VRA without the most frequent verbs for 3sg subjects are presented in Table 45 and the comparison between the two analyses is presented in Table 46.

The results for the analysis of 3sg subjects reveal that 3sg subjects do not appear to be as strongly affected by high frequency verbs as the other two persons. This is rightly

observed in the rate of pronominal expression. There is barely a change in the rate of pronominal expression when the high frequency verbs are excluded from the analysis (48% with all verbs, 47% when we exclude the high frequency verbs). This suggests that the patterning these verbs are forming with pronominal subjects is not impacting these subjects in the same way.

Turning now to the differences between the two analyses, we can observe that there is major change in the way the factor groups are organized in these data. Initially, modal was the strongest factor group in significantly conditioning the realization of the variant. In the analysis without the most frequent verbs, on the other hand, modal is no longer a significant predictor of the realization of pronominal 3sg subjects, and, although there is minimal difference, the direction of effect is reversed. This finding helps us support our hypothesis that these high frequency verb forms are forming their own patterns of pronominal expression and that these individual patterns are swaying the distribution of the data.

A second observation obtained from this analysis concerns discourse continuity. This factor group is now the strongest predictor of pronominal subjects in the data and it shows the same patterning observed in the earlier analysis, namely that more continuous contexts disfavor the occurrence of pronominal subjects, whereas less continuous contexts offer the loci for the realization of pronominal subjects. This may be due to **MODAL** no longer being selected as significant, so it is not so much that **DISCOURSE CONTINUITY** has become more significant, but **MODAL** has lost its significance.

Table 45. Multivariate Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 3sg subjects without the most frequent verbs

Total *N*	1298			
% *expressed*	46.7			
Corrected *Mean*	.466			
	Probability	**% expressed**	**N**	**% data**
Discourse continuity				
Diff Subj	.58	54.4	539	41.5
Same Subj & Diff TAM	.46	42.9	331	25.5
Same Subj & Same TAM	.43	40.0	428	33.0
Range 15				
Polarity				
Affirmative	.51	47.6	1148	88.4
Negative	.42	39.3	150	11.6
Range 09				
Clause type				
Subordinate	.57	52.9	189	14.6
Main	.49	45.6	1109	85.4
Range 08				
Verb class				
Possession	[.51]	47.5	49	03.8
Other	[.51]	46.9	1050	80.9
Cognition	[.46]	43.2	125	09.6
Speech	[.44]	40.5	74	05.7
Range *n.s.*				
TAM				
Imperfect	[.54]	49.7	157	13.0
Present	[.50]	46.0	771	63.9
Preterit	[.49]	45.7	278	23.1
Range *n.s.*				
Morphological irregularity				
Regular	[.50]	47.1	1139	87.8
Irregular	[.47]	43.4	159	12.2
Range *n.s.*				
Modal				
Present	[.51]	47.8	249	19.2
Absent	[.50]	46.4	1049	80.8
Range *n.s.*				

Total Chi-square = 178.6778; Chi-square/cell = 0.8759; Log likelihood = -879.437
Factor groups not selected as significant: Modal, Morphological irregularity, TAM, and Verb class.

Finally, a third finding borne out from this analysis is the selection of polarity as a significant predictor of pronominal subjects. The pattern observe with these subjects is the same observed with the other two and the same when all verbs are included, i.e. affirmative statements favor the realization of pronominal subjects, and negative statements disfavor the occurrence of pronominal subjects. The effect of this factor group is not as clear to understand. I argued earlier for 2sg and 3sg subjects that there seemed to be a strong link between some predicates and negation markers in co-occurring without pronominal subjects. However, this does not seem to be the case with these subjects, instead what may be happening is that these forms with negation markers are so entrenched in occurring in contexts without pronominal subjects that they are slower in adopting the variant choice, thus their disfavoring in these data.

A final remark concerning the results observed for 3sg subjects is directed to morphological irregularity, which is the only factor group in these data to not be selected in any of the analyses, as was the case for the analysis that included all verbs, and the direction of effect (though the difference is minimal) remains the same. This suggests again that 3sg subjects are not as prone to lexical effects as the other two persons. While the data for the analysis shows that there are blatant differences between these two datasets, they are not lexically motivated as the differences observed for 1sg and 2sg subjects. Again, this is conjectured to the result of the category of 3sg being populated by a lot more verbs than that of the other two persons.

Table 46. Comparison of Multivariate Analysis of the factor groups that contribute to a statistically significant result on the conditioning of pronominal expression of 3sg subjects.

		3sg – All tokens		3sg – High frequency verbs excluded
Total *N*		2930		1298
% *expressed*		47.7		46.7
Corrected *Mean*		.475		.466
		Prob.		**Prob.**
Verb class				
Perception		[.59]		
Possession		[.56]		[.51]
Cognition		[.51]		
Other		[.50]		[.51]
Speech		[.48]		[.44]
Relational		[.45]		[.46]
Range	*n.s.*		*n.s.*	
Clause type				
Subordinate		.60		.57
Main		.48		.49
Range	12		08(3)	
Discourse continuity				
Diff Subj		.56		.58
Same Subj & Diff TAM		.48		.46
Same Subj & Same TAM		.43		.43
Range	13		15(1)	
TAM				
Preterit		.53		[.49]
Imperfect		.53		[.54]
Present		.48		[.50]
Range	05		*n.s.*	
Polarity				
Affirmative		[.51]		.51
Negative		[.45]		.42
Range	*n.s.*		09(2)	
Modal				
Absent		.53		[.50]
Present		.39		[.51]
Range	14		*n.s.*	
Morphological irregularity				
Regular		[.50]		[.50]
Irregular		[.50]		[.47]
Range	*n.s.*		*n.s.*	

While **TAM** is no longer selected as significant in this analysis, it is worth noting that the **PRETERIT** most disfavors pronominal subjects, whereas before it most favored. This is important because both past tenses, with all verbs, favored the realization of pronominal subjects, which followed the prediction of their occurring with pronominal subjects because they are the more pragmatically marked tenses. With this analysis in mind, what we see is that it is really the **IMPERFECT** that affects pronominal subject realization.

One last note on **TAM** concerns its magnitude of effect from the analysis with all verbs to the analysis without the high frequency verbs. In the latter analysis, this factor group is not selected as significant, and this is partly due to the correlation between the highly frequent predicates excluded from the analysis and the **PRESENT** TAM.

6.5 Discussion and Summary

In this chapter I have outlined the results for VRAs of 1sg, 2sg, and 3sg subjects without the most frequently occurring verbs in each dataset. In each analysis I excluded all verbs whose token count consisted of one percent or more of the total number of verb tokens in the dataset, and the findings that were borne out from these protocols supported the hypotheses established earlier in this work.

Three important findings are obtained by these analyses. Firstly, for 1sg and 2sg subjects there was a considerable increase in the rates of pronominal expression in the data with the most frequently occurring verbs removed (from 64.7% to 71.8% for 1sg subjects and from 53.4% to 67.2% for 2sg subjects). An implication of this is the possibility that the data without these highly frequent verbs is behaving "normally" in the sense that it may not contain any abnormal patterns that significantly deviate from the expected variation within the grammar of speakers.

A second finding achieved by these separate analyses is in the variation in the strength of effect of different factor groups in each of the persons. We had the opportunity to observe factor groups that were strong predictors of pronominal expression in one analysis then not be selected as significant in the second; factors that had a somewhat weaker effect in one analysis became stronger; factors that were not present in the first analysis emerged in the second. Finally, despite all the changes in significance among the different factor groups, the changes in direction of effect observed are more significant to the argument that these persons are behaving differently. And most importantly, none of the persons patterned similarly according to the factors that were selected as significant, with the exception of clause type that remained as the only factor group to condition variant choice in all analyses in this study.

The implication of these changes in strength of effect, factor groups being significant in one analysis but not in the other, and changing directions of effect, suggest that the grammaticization of pronominal subjects in BP is still very much under way and far from being complete. This combination of findings provides some support for the conceptual premise that there is more variability in the way pronominal expression manifests in the language than many researchers would like to admit. However, such variability does not take place in a random and disordered way, there is still some possibility of making sense of it by breaking apart the different forms in the language that seem to be behaving independently and those forms that seem to be converging into a more organized, so to speak, patterning. This is what this study is attempting to do, namely, dissect the different contexts and forms that contribute to the spread of pronominal expression and to the retention of unexpression.

Finally, the third finding attained by these analyses is reflected in the changes in some of the hierarchy of constraints for some of the factor groups selected as significant. These changes are most noticeable in the factor group TAM with 2sg subjects where both past tenses are favoring pronominal expression as opposed to only the imperfect in the earlier analysis. Moreover, the present tense, which favored pronominal subjects in the earlier analysis, now disfavors the occurrence of such forms. This reorganization of these tenses with 2sg subjects fall more in accordance with what would be expected of them in terms of their patterning with pronominal subjects than the pattern observed in the analysis with all the verbs. Because the present tense is the more frequent tense in the language, it is expected that it will take on the pattern of pronominal expression more slowly than the other less frequent tenses, which, because of their low frequency, are more likely to be subsumed by the emerging pattern.

The present results are significant in at least major two respects. It shows that each person is behaving distinctly from the others at several levels even when the predictors used in the analysis are kept constant. This finding, thus, implicates any further analysis of pronominal expression in BP, and possibly in other languages as well, to examine each person in their own domain without trying to analyze all as one whole. The analysis of all persons combined, at least in these data, did not reveal a solid understanding of the underlying structure or variation of pronominal subjects in BP.

The second significance of these findings requires a further breakdown of the data to analyses that exclude predicates of high frequency. I hope to have shown in the analyses presented in this chapter that there is a pronounced difference between the data with the most highly-occurring verbs and the data without them. This finding was echoed across the three

different persons with varied effects throughout the different factor groups included in the study. This can be attributed to the fact that forms of high frequency tend to behave differently and more individually than the rest of the language (Barddal, 2006; Barddal & Eythórsson, 2003; Barddal, et al., 2011; Hay, 2001; Hay & Baayen, 2002, 2005), thus these verbs of high frequency that were excluded from the analyses represent forms with their own individual patterning with pronominal subjects, and thus they cannot be taken, in any way, to represent the way the language is manifesting its distribution of pronominal subjects in the datasets analyzed.

7 CONSTRUCTION EFFECTS

In the previous chapter I outlined the results for the VRAs without the most frequent verbs for each person, in order to examine the conditioning of variant choice when the seven predictors, verb class, clause type, TAM, modal, morphological irregularity, discourse continuity, and polarity are considered together. It was observed that there is a resounding difference between these analyses and the ones presented in Chapter 5. Thus, in this chapter, I will examine these high frequency verbs and other contexts that I consider to contribute to the overall pattern of subject expression in these data.

7.1 Introduction

One of the crucial premises of Variationist theory has its core in the circumscription of the variable context or the envelope of variation, i.e., the lieu in which all the variants under consideration may occur and overlie one another. Tagliamonte (2006), following Guy (1988), posits that contexts that occur at extremes (e.g., at 95% or at 5%) should not be included in any variable rule analysis for these contexts do not behave in the same way as the rest of the data in relation to the variable. They are deemed categorical in nature and should not be analyzed. Otheguy et al. (2007) go farther than Tagliamonte in that they establish the inclusion or exclusion of certain contexts based on their low or high variability rather than "one between absolutely variable and absolutely invariable environments" (Otheguy, et al., 2007, p. 776). Thus, the grounds for what can be included or not within the envelope of variation must be established by the degree of variability of the context. So, however this context is established, it is standard procedure to exclude tokens outside this envelope from further quantitative studies (Aaron, 2006; Poplack & Tagliamonte, 1991). In sum, scholars

agree that we need to remove contexts that show no variation or very little variability. Nevertheless, the exclusion of these contexts begs the question of their role in contributing to the distribution of pronominal subjects in these data. And this is the premise of this chapter, to investigate and analyze these contexts and argue, not for their inclusion in statistical analyses, but for inclusion in the overall analyses of such a complex pattern as subject expression.

In the study of pronominal subjects being conducted here I excluded several contexts of categorical occurrence or non-occurrence of pronominal subjects and in Chapter 6 I argued for the exclusion of high frequency verbs as well. Thus, here I will focus on two portions of the data, namely the contexts excluded from the envelope of variation, most specifically those contexts of high frequency that may contribute to the spread of pronominal subjects or the retention of their counter-variant. The second portion to be examined will be the highly frequent verbs that were excluded from the analyses in the previous chapters. Some of these verbs form highly entrenched constructions that favor one or the other pattern, while other are just frequent in their occurrence, but because of their frequency they may be considered to form constructions as well even though their patterning may not be as entrenched as the patterning of other forms. In short, what we will see in the analysis of these constructions is that subject realization forms part of them, while in others subjects are not part of the construction.

Constructions have been defined in many ways throughout the literature, but we are adopting the notion that they are pairings of form and meaning (Bybee, 2010; Bybee & Cacoullos, 2009; Croft, 2001; Goldberg, 2006)(Torres Cacoullos, 2006; Torres Cacoullos & Travis, 2010; Torres Cacoullos & Walker, 2009) that may or may not include pragmatic

value. Scholars have concurred that constructions can be anything from "mono-morphemic words, to complex words, to idioms, all the way up to general configurations such as the 'passive construction'" (Bybee, 2010, p. 76). In this study, constructions are to be understood as pairings of verbs with tense and specific subjects (1sg, 2sg, 3sg) of high frequency[40]. This conception of a construction can account not only forms that have an entrenched pragmatic function in discourse, but also for those verbs that have become so frequently used together without necessarily having attained a function or new meaning, as well as those forms that have not become generalized or bleached of meaning. In short, constructions are viewed here as sequences of frequently co-occurring forms. Take examples (46) and (47) to illustrate this definition.

(46) *só sei que tem isso*
'all (I) know is that they have it'
(I9: 576)

(47) mas eu já tenho esse horário em função...
'but *I already have* this schedule because...'
(C7: 384)

The first example illustrates the construction *sei* '(I) know', which occurs most of time without pronominal subjects (cf. discussion later in this chapter), while the second example illustrates a highly frequent verb to occur with 1sg subjects that does not form a construction on its. However, these forms tend to influence the pattern of expression in similar ways, namely, each forms its own individual patterning with pronominal subjects, which crystalize contributing thus to the change toward obligatory subject expression. The

[40] Please recall that high frequency as adopted in this study refers to the verbs that accounted for at least one percent of the data of each person independently, i.e. verbs that attained such measure among 1sg subjects may not emerge with 2sg and 3sg subjects.

opposite effect is also true in that certain constructions and highly frequent verbs that most frequently occur without pronominal subjects contribute to the retention of this pattern.

To understand how the high-frequency verbs are patterning with each person, I will divide them per person in the next three subsections. In each subsection I will describe the most frequent verbs, and any patterns of predicate and tense observed for each person will also be discussed. The main objective of these subsections is to illustrate how these high frequency verbs are affecting the way pronominal expression is manifested in BP. The fourth and last subsection of this section describes the constructions that were excluded from the envelope of variation and how they correlate with pronominal expression.

7.2 1sg Subjects

The most frequently occurring verbs with 1sg subjects (those that account for 1% or more of the total number of 1sg verbs) account for 66% of all 1sg clauses. The verbs are documented in Figure 10 where we can see their distribution according to pronominal expression. However, the table below clearly illustrates that not all predicates show the same rates of frequency. Indeed, we see that there is an obvious demarcation between these subjects in terms of their frequency, forming five distinct groups of verbs depending on how they are divided in terms of their frequency of occurrence.

It is worth noting that while these verbs are being examined according to their frequency here, they all favor pronominal expression throughout. This is most enlightening because, overall, the forms that are used more often in the language with 1sg subjects are forms that occur most frequently with pronominal subjects given some exceptions that will be discussed below.

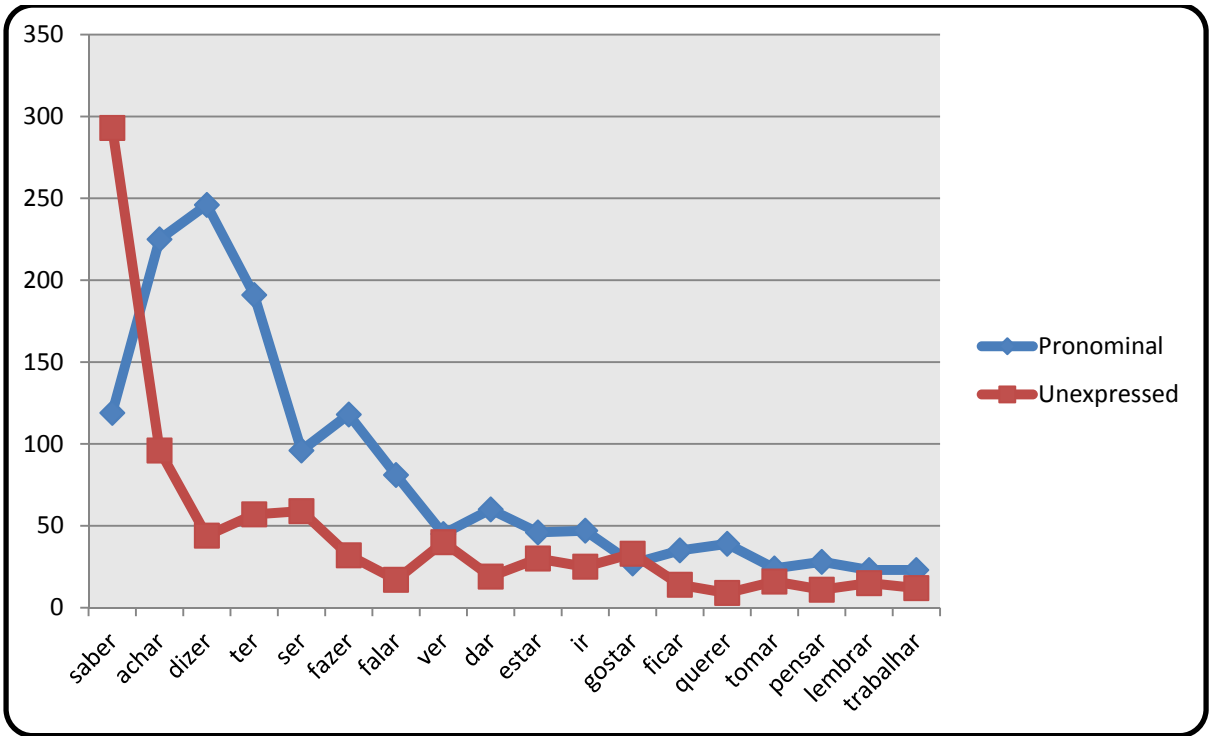

Figure 10. *Verb types representing 1% or more of 1sg data.*

The most frequently-occurring predicate with 1sg subjects is the verb *saber* 'to know', which amounts to nearly 20% of this set of the most frequent verbs with 1sg subjects. This verb is also that which most favors unexpressed subjects in these data, with a rate of non-expression of 71% (119/168).

The second segment of high frequency verbs consists of the verbs *dizer* 'to say' (85% pronominal, 246/290), *ter* 'to have' (77% pronominal, 191/248), and *achar* 'to think' (70% pronominal, 225/321). These three verbs account for 37.44% of the data of high frequency verbs occurring with 1sg subjects where each accounts for one third of this total, and together they account for 25% of all instances of predicates with 1sg subjects. These three verbs show a strong favoring of pronominal subjects; nearly three fourths of these verbs occur with pronominal subjects. This suggests that these verbs are patterning similarly in terms of their distribution with pronominal subjects, but differently from *saber* 'to know'.

The next set of high frequency verbs accounts for 13% of the data divided almost evenly between two predicates, *fazer* 'to do/make' (79% pronominal, 118/150) and *ser* 'to be' (62% pronominal, 96/155).These verbs favor pronominal expression with the latter favoring them more strongly than the former.

The fourth group of predicate consists of six different verbs amounting to approximately 20% of all the high frequency verbs excluded from the analysis. This is a very heterogeneous category of verbs and it only comprises four verbs that strongly favor pronominal expression, i.e. *falar* 'to speak' (83% pronominal, 81/98) and *dar* 'to give' (76% pronominal, 60/79) and that show a preference for the realization of pronominal subjects, namely *estar* 'to be' (61% pronominal, 46/76) and *ir* 'to go' (65% pronominal, 47/72).

The last group of verbs to highly occur with 1sg subjects consists of six verbs as well, namely *ficar* 'to stay', *querer* 'to want', *tomar* 'to take', *pensar* 'to think', and *lembrar* 'to remember', amounting to approximately 11% of the data. This group is very homogeneous in terms of its patterning with pronominal expression with all verbs occurring with pronominal subjects at at rate of 60% and above. The demarcation described in these paragraphs is illustrated in Table 47.

The patterns of frequency observed in the table above suggest that the verb *saber* 'to know' is a clear example of the conserving effect of high-frequency forms (Bybee, 2002, 2010; Bybee & Cacoullos, 2009). These forms are strengthened every time they are used, which facilitates their access for future use and impedes the advancement of changes within their structures (Bybee, 2010, p. 24). This is exactly the case with this verb when it occurs with 1sg subjects, more specifically without the realization of the pronoun *eu* 'I', which takes place almost three fourths of the time in these data.

Table 47. Rates of expression of most frequenct verbs with 1sg subjects.

		Pronominal	
		N	***%***
Frequency set I	*saber* 'to know'	119	29
Frequency set II	*achar* 'to think'	225	77
	dizer 'to say'	246	
	ter 'to have'	191	
Frequency set III	*ser* 'to be'	96	70
	fazer 'to do/make'	118	
Frequency set IV	*falar* 'to speak'	81	65
	ver 'to see'	45	
	dar 'to give'	60	
	estar 'to be'	46	
	ir 'to go'	47	
	gostar 'to like'	27	
Frequency set V	*ficar* 'to stay'	35	69
	querer 'to want'	39	
	tomar 'to take'	24	
	pensar 'to think'	28	
	lembrar 'to remember'	23	
	trabalhar 'to work'	23	
	Total	**1473**	64

The high frequency of the verb *saber* 'to know' with unexpressed 1sg subjects cannot be the only reason why unexpression still lingers with 1sg subjects; however, these figures are very suggestive of the effect this form has over the distribution of expression in the language, and the examination of these high frequency verbs has demonstrated that. We can see here that this verb is behaving noticeably differently from the other high frequency verbs and the other verbs from its semantic class, cognition. Three other verbs of cognition have achieved the status of highly frequent verbs to co-occur with 1sg subjects in BP, a finding that has been reported for both American English (Scheibman, 2002) and for BP as well (Silveira, 2007), namely *achar* 'to think', *lembrar* 'to remember', and *pensar* 'to think'. These three cognition predicates strongly favor expressed subjects (69% pronominal,

276/398). In short, the verb *saber* 'to know' has formed its own structural behavior in relation to pronominal expression, and such behavior has implications that affect the pattern of subject expression with 1sg subjects.

A more detailed analysis of this verb form reveals that three constructions with the verb *saber* 'to know' and 1sg subjects in the present tense are contributing to the conserving effect observed in these data. The first construction is *sei* '(I) know$_{\text{-PRESENT}}$' which represents 32% of the tokens of this verb with 1sg subjects (131/412). The other two constructions occur in the negative, namely *não sei* '(I) don't know$_{\text{-PRESENT}}$' (N= 24) and *num sei* '(I) don't know$_{\text{-PRESENT}}$' (N=73), with their combined count amounting to 24% of all occurrences of the verb *saber* 'to know' in these data (97/412). Examples of each construction are illustrated in (48), (49), and (50) respectively.

(48) *A: Depois eu tenho também dicionário da Bíblia... que até um... um amigo meu o pastor S. de Cuba que me deu...*
... aquele... que eu entrevistei
B: Sei.
A: Que eu fui fazer pesquisa.
'A: Besides I also have the Bible dictionary ... which ... a friend of mine, pastor S. Cuba gave it to me...
... that one .. who I interviewed
B: (I) know
A: When I was researching'
(inq. 33:732)

(49) *num sei… ele cantou do teatro*
'(I) don't know... he sang from the theater'
(C48: 1019)

(50) *não sei… tenho coragem de voltar aquela vida antiga*
'(I) don't know... (I) have the courage to get back to that life'
(C116: 71)

These three constructions together account for 56% of all tokens of *saber* 'to know' in these data, suggesting that they may play a decisive role in the distribution of pronominal subjects with 1sg subjects. Studies that emphasize the effects of frequency in lexical retrieval

and representation have stressed the importance of the most frequent member of the category, which is alleged to draw the other members of the category or the exemplar cloud toward it, thus, through analogy, the attracted members will tend to behave similarly to the patterns manifested in the most frequent member of the category. The three constructions exemplified in (48), (49), and (50) are indeed instances of the most frequent members of a category, here the realization of the verb *saber* 'to know' with 1sg subjects. They draw the other possible forms of *saber* 'to know' to be realized in a similar fashion, that is, without the presence of a pronoun.

In contrast to the patterning observed with the verb *saber* 'to know', the remaining four sets of high frequency verbs exhibit a similar patterning for the realization of pronominal 1sg subjects in these data, with the second set of verbs showing a stronger preference. In the remainder of this section I will focus on the possible constructions that emerge from this set of verbs.

The first predicate from this group is *achar* 'to think', which has a rate of pronominal expression of 70%. The most frequent construction to emerge in these data is *eu acho* 'I think$_{\text{-PRESENT}}$' illustrated in example (51), which accountd for 49% of all tokens of *achar* 'to think'.

(51) *A Sulamita me disse que mil e quinhentos dá.Mas num sei. Eu acho que dá porque são dezessete pessoas, né?*
'Sulamita said that a thousand and five hundred would be enough. But I don't know. I think it would, after all there are seventeen people, right?'
(Inq. 7:685)

The second predicate in this group is *dizer* 'to say', which shows a rate of expression of 85%. The most frequent construction with this verb is *eu digo* 'I say$_{\text{-PRESENT}}$', which accounts for 35% of all occurrences of *dizer* 'to say' in these data as opposed to *digo* '(I)

say', which only accounts for 4% of all occurrences of *dizer* 'to say'. An example of *eu digo* 'I say$_{\text{-PRESENT}}$' is illustrated in (52) below.

(52) *como <u>eu digo</u> a você…*
'as <u>I say</u> to you'

(I5: 687)

The third predicate in this group is *ter* 'to have', which shows a rate of expression of 77%. The construction *eu tenho* 'I have' accounts for 38% of all occurrences of this predicate with 1sg subjects as opposed to *tenho* '(I) have', which only occur 11% of the time, in other words, three times less often than *eu tenho* 'I have', which is illustrated in example (53) below.

(53) *como profissional <u>eu tenho</u> apenas quatro anos de experiência*
'as a professional <u>I have</u> just four years of experience'

(I21: 284)

One final note on these constructions refers to their variability in terms of pronominal expression. While *saber* 'to know' shows a strong favoring for unexpressed subjects and the other three predicates favor pronominal expression, they do not demonstrate a categorical behavior with the forms they prefer. What we see with them is a tendency to occur more frequently with one form or the other. Thus, it is hypothesized that over time *saber* 'to know' will likely show greater rates of pronominal expression, with the exception of the constructions discussed here, i.e. *não/num sei* '(I) don't know', which account for 47% of all tokens of *saber* 'to know' and they may become more fossilized as relics of non-expression.

7.3 2sg Subjects

The most frequently occurring predicates with 2sg subjects account for 73% of all tokens extracted for these analyses. The distribution of these predicates with pronominal subjects is given in Figure 11. While the verb *saber* 'to know' disfavors pronominal expression with 1sg

subjects, all other verbs to occur with this subject homogeneously favor pronominal subjects. This is not what is found for 2sg subjects, the most frequently occurring predicates with 2sg subjects do not demonstrate such homogeneity.

As the figure below reveals, there are five predicates that show the highest frequency of this group of verbs and the most variability, namely *saber* 'to know', *entender* 'to understand', *ter* 'to have', *ver* 'to see' and *olhar* 'to look'. In the remainder of this section I will discuss the constructions that emerge with these predicates and how they contribute to the overall patterning of pronominal expression with 2sg subjects.

The patterning of these predicates when compared to the remaining predicates raises the question of why the rest of the predicates were included in the analysis. As was stated earlier, preliminary analyses were conducted to test whether the presence or absence of these predicates would result in significant changes in the log likelihoods of the analyses, or the magnitude of effect of each factor group, or the hierarchy of constraint within each factor group, as well changes in direction of effect. None of these factors emerged as differences from separate analyses, thus it does not matter statistically whether these verbs are included in the analysis or not. So, the choice to retain these verbs in these analyses is both linguistic and methodological. As this is one of the first studies to suggest this type of analysis (cf. Goddard, 2005), the approach is experimental, and more future studies should illuminate this methodology on whether it should be maintained or reformulated.

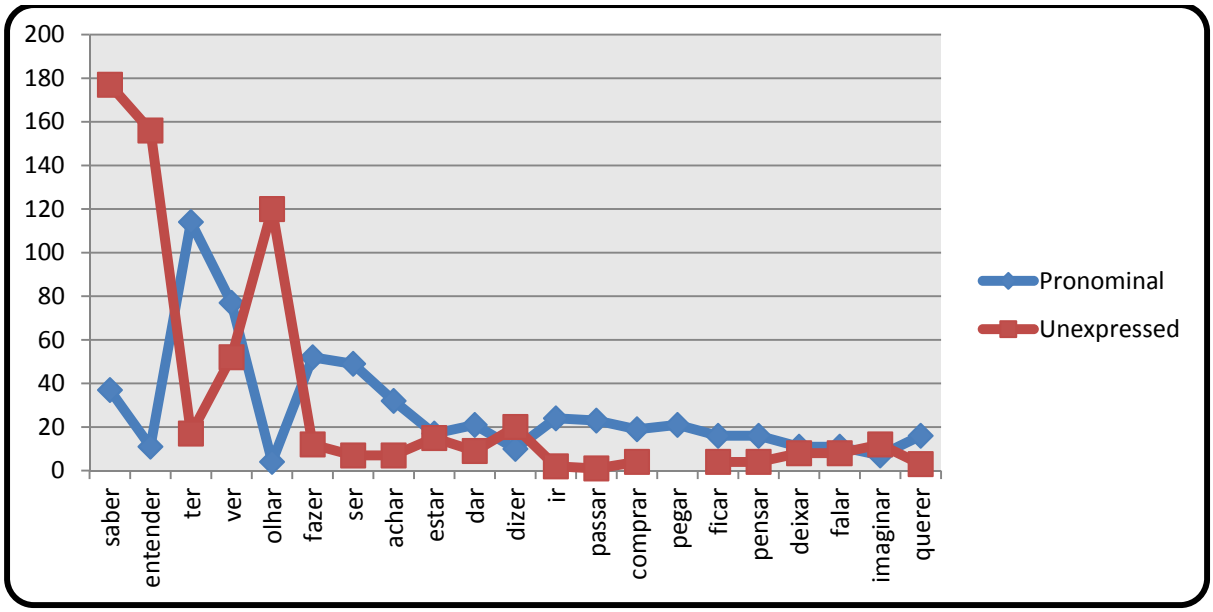

Figure 11. Distribution of most frequently occurring verbs with 2sg subjects according to pronominal expression.

Similarly to 1sg subjects, the most frequently occurring predicate with 2sg subjects is the verbs *saber* 'to know'. This verb alone accounts for 13% of all 2sg subjects. With this verb we see the emergence of one construction from the combination of 2sg subjects, the verb *saber* 'to know', and the present tense, namely *sabe* '(you) know$_{\text{-PRESENT}}$' which accounts for 9% of all tokens of 2sg subjects. This construction is illustrated in (54). The remaining forms of the verb *saber* 'to know' follow the same patterning observed in *sabe* '(you) know$_{\text{-PRESENT}}$' in that they tend to be realized without pronominal subjects.

(54) *todo MUNdo consegue fazer o trabalho que a gente faz, tá entendendo?... ma/ num é:: num é bem assim... sabe?...*
'everybody can do the work that we do, (you) understand? Bu-
(it) is not (it) is not like this…(you) know?'
(I106:24)

The second most frequent predicate to occur with 2sg subjects in these data is *entender* 'to understand', which accounts for ten percent of all occurrences of 2sg subjects. Three constructions are observed with this predicate in our data, namely *entende* '(you)

understand$_{\text{-PRESENT}}$' illustrated in (55), *tá entendendo* 'do (you) understand$_{\text{-PRESENT PROGRESSIVE}}$' illustrated in (54), and *entendeu* '(you) understand$_{\text{-PRETERIT}}$' illustrated in (56).

(55) *já tinha saído a maioria do pessoal, <u>entende</u>, eu sonho com essa velha...*
'everyone had already left, *(you) understand*, I dream about this old woman...'
(I13: 461)

(56) *eles não são professores da Universidade, <u>entendeu</u>?*
'they are not professores at this university, <u>(you) understood</u>?
(I27: 1225)

The construction *tá entendendo* 'do (you) understand$_{\text{-PRESENT PROGRESSIVE}}$', illustrated in example (54), is the most frequent construction to occur with 2sg subjects and this predicate, accounting for 43% (73/167) of all occurrences of this verb with 2sg subjects. The second most frequent construction is *entendeu* '(you) understand$_{\text{-PRETERIT}}$', illustrated in (56), accouting for 23% of all tokens of this predicate (39/167). Finally, the third most frequent construction with *entender* 'to understand', *entende* '(you) understand$_{\text{-PRESENT}}$', illustrated in example (55), accounts for 19% of all occurrences of *entender* 'to understand' with 2sg subjects (31/167). These three constructions amount to a total of 91% of all tokes of *entender* 'to understand' and 9% of all 2sg singular tokens. Finally, these constructions account for 93% of all instances of non-expression with this verb.

The verb *ter* 'to have' is the third most frequently co-occurring verb with 2sg subjects in these data accounting for eight percent of all 2sg tokens. The most frequent form within the occurrences of this predicate is with pronominal subjects in the present tense, which account for 87% of the occurrences of *ter* 'to have', and 90% (57/63) of them occurred with a pronominal subject. An example of this form is illustrated in (57).

(57) *<u>você tem</u> o material do ano anterior…*
'<u>you have</u> the material from the previous year...'
(C16: 347)

The remaining two verbs to highly occur with 2sg subjects exhibit opposing preferences in their distribution of pronominal subjects. The predicate *ver* 'to see' occurs with 2sg subjects eight percent of the times, of which 60% of these occurrences are realized with pronominal subjects in the present tense. An example of this pattern can be seen in (58) below. The predicate *olhar* 'to look', on the other hand, strongly disfavor the occurrence of pronominal subjects, occurring with them only 3% of the time. This predicate accounts for seven percent of all tokens of 2sg subjects, of which 73% are tokens of the construction *olha* '(you) look$_{\text{-PRESENT}}$'. This construction is illustrated in example (59).

(58) *Você vê que o exército é uma escola.*
'You see that the army is a school.'
(I5: 127)

(59) *eu conversei agora há pouco tempo com um rapaz ele teve nos Estados Unidos "rapaz olha é imPREssionante como lá nos Estados Unido tão ensinando o espanhol agora"*
'i talked a while ago with a guy he was in the USA "look (it) is impressive that in the US they are teaching Spanish now"'
(C47: 413)

There are two implications that can be extracted from these findings for 2sg subjects. Firstly, unlike 1sg subjects, the patterns observed for 2sg subjects do not show homogeneous behavior by most of the verbal forms, at least not for the highest frequency ones. The five most frequent predicates discussed in this section show very different patterns of pronominal expression with 2sg subjects i.e. *olhar* 'to look' with 97% of non-expression, *entender* 'to understand' with 93% of non-expression, and *saber* 'to know' with 83% of non-expression, as opposed to *ter* 'to have' with 90% of expression and *ver* with 60% of pronominal expression. This leads to the second implication of these findings that the careful examination of patterns of verbs and subjects reveal patterns of syntactic behavior that defy the generalizations abstracted from statistical analyses or linguistic models, which, vis-à-vis

1sg subjects, appear to govern the way pronominal expression is distributed through the language.

7.4 3sg Subjects

Third person animate singular subjects are the second most frequent subjects in the data analyzed in this study. Following the pattern observed for 1sg and 2sg subjects, 3sg subjects occur frequently with a small number of verbal types, illustrated in Figure 12. However, this tendency is not as great as with the other two persons. While 1sg and 2sg subjects occur with a small number of verbal types 66% and 73% of the time, respectively, 3sg subjects only occur 57% of the time with verbal forms that have more than one percent of token realization with these subjects.

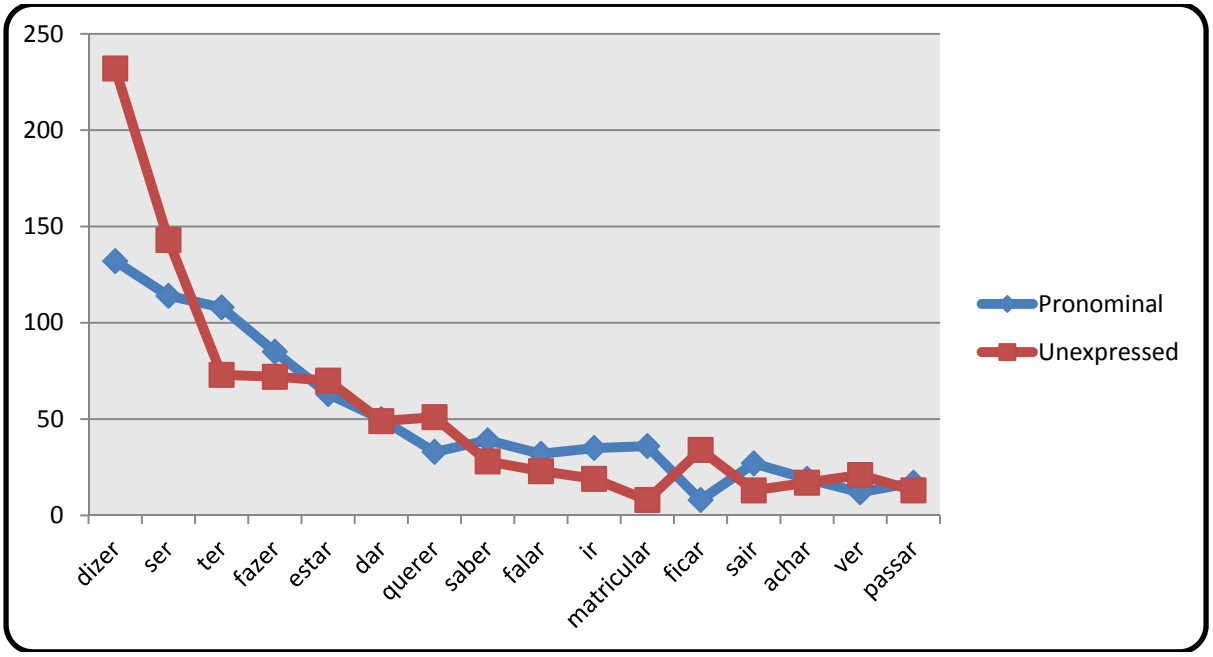

Figure 12. Distribution of high frequency verbs with 3sg subjects according to pronominal expression.

Four predicates appear to be forming patterns that may correspond to the overall distribution of pronominal expression in the data for 3sg subjects. These verbs are *dizer* 'to

say', *ser* 'to be', *ter* 'to have, and *fazer* 'to do/make', with their rates of expression illustrated in Figure 13. As can be noted in the figure below, the two most frequently-occurring verbs occur more often with unexpressed subjects (though much more markedly for *dizer*[41] than for *ser*), whereas the next two most frequent verbs occur slightly more often with expressed subjects. Most importantly, none of these predicates form noticeable constructions as the ones observed for 1sg and 2sg subjects. Rather, they are just the most frequently occurring predicates with 3sg subjects.

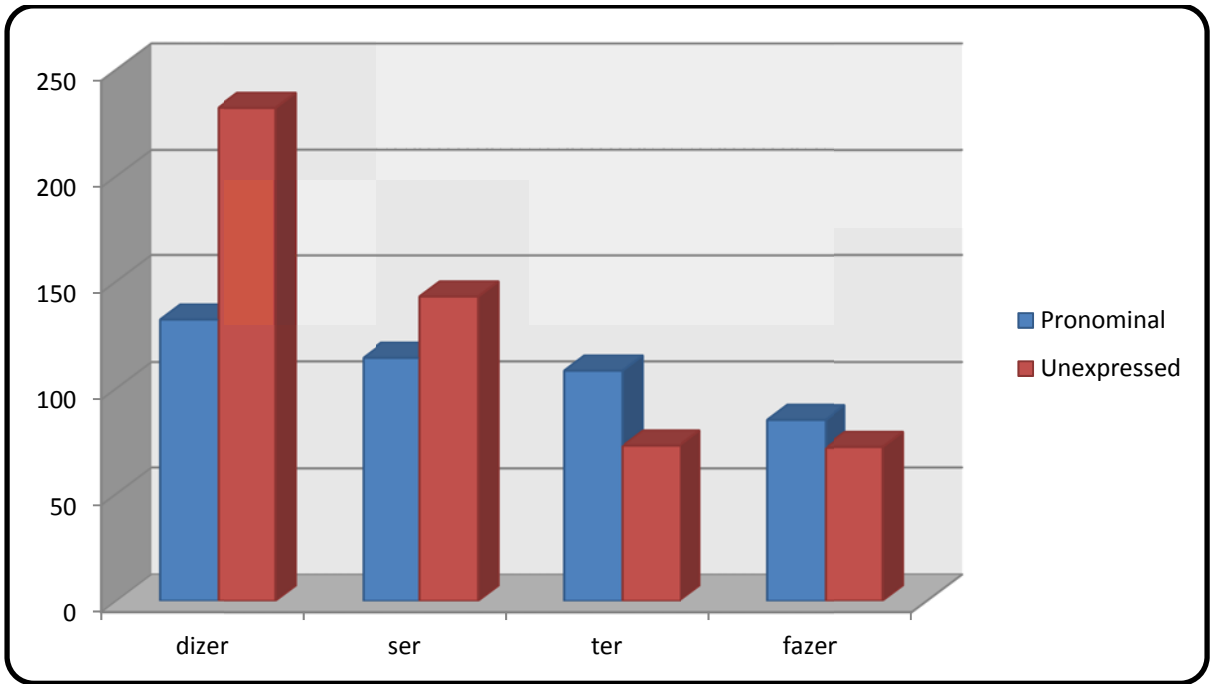

Figure 13. Distribution of four most frequent verbs to occur with 3sg subjects according to their rates of pronominal expression.

The patterns observed from the figure above suggests that a lot of discourse produced with 3sg subjects consists of the schema illustrated in (60) (M. A. K. Halliday, 1994) where

[41] The pattern observed for this verb with 3sg subjects is exactly the opposite of the pattern observed for 1sg subjects, so this is another reason why collapsing these predicates does not offer a clear picture of the overall distribution of pronominal expression in the language.

(a) illustrates the pattern of a speech predicate, (b) and (c) the pattern of relational predicates[42], and (d) the pattern of a verb of doing.

(60) Schemas for 3sg subjects and predicates

(a) *(x) diz y* '(x) says y'
A gente fala "como é bom te ver hoje, Davi" e ela diz "hoje Davi" ela só fala o final da frase.
'We say "it is good to see you today, Davi" and she says "today Davi" she only speaks the end of the phrase.'
(I21: 680)

(b) *(x) é y* '(x) is y'
Inf. 1: Eu tava vendo a qualidade desse curso
Inf. 2: é muito bom
'Inf. 1: I was loking at the quality of that course
Inf. 2: (it) is very good.'
(C16: 155)

(c) *x tem y* 'x has y'
Inf. 1 - inclusive na própria União Soviética...
quer dizer o pessoal...
que fazia parte do partido...
ele tinha os privilégio de comprar no supermercado.
'Inf. 1 - even in the Soviet Union...
I mean the people…
who was part of the political party…
had the privilege of buying at a supermarket.'
(C30: 245-248)

(d) *x faz y* 'x does/makes y'
Ele coloca o teu sistema começando aqui... né? a maioria dos livro... e ele faz... com que ele tivesse aqui dois eles.
'He places the system starting here... right? The majority of books … and he makes it look like there were two of them here.'
(L54: 521-527)

[42] While throughout this analysis the predicate *ter* 'to have' has been treated as a **POSSESSION** predicate, its syntactic behavior is more in line with the behavior of **RELATIONAL** verbs where the predicate established a relationship between *x* and *y*.

These schemas depicted in (60) also suggest that these structures form topic-comment statements (Lambrecht, 1994) where something, *y*, is said about *x*. As such, the pattern of pronominal expression observed in (a) and (b) would be the predictable pattern since these statements tend to be used in discourse after the *x* has been introduced (Lambrecht, 1994, 2001), thus being old information, and more likely to be left unexpressed. However, the pattern observed in (c) and (d) does not conform to this hypothesis, suggesting that even within these high frequency verbs, the pattern of expression is creeping in.

7.5 Other Constructions

Among the contexts excluded during the culling of the datasets were a group of constructions that categorically disfavor pronominal subjects. These constructions are documented in Table 48 and they correspond to 15% (2887/18810) of all the tokens extracted for this study of pronominal expression.

Table 48. Constructions that categorically occur without pronominal subjects.

Constructions				
	N	%		
Existential constructions	640	22.17	2887	15%
É	311	10.77		
Era	832	28.82		
comé	226	7.83		
é que	189	6.55		
quer dizer	165	5.72		
entende?	125	4.33		
viu?	219	7.59		
Seja	98	3.39		
será	34	1.18		
Other	48	1.66		

Note that constructions occur across the board, not only with human subjects as the ones presented in the previous subsections (e.g., *não sei* '(I) don't know), but with non-

human subjects as the ones that are going to be dealt with in this section. The first construction to be observed is the construction formed with existential predicates. Two forms, illustrated in (61) and (62), are *há* 'there is/are' and *tem* 'have$_{-3SG\text{-}PRESENT}$', are the most frequent ones in the category. They are always followed by an argument, but this argument is more likely to be a topic then a subject since these types of constructions are characteristically thetic in nature (Lambrecht, 1994).

(61) *Eu fumava três cigarros por dia e há pessoas que diz que três cigarros por dia não faz mal a ninguém.*
'I smoked three cigarretes a day and there are people who say that three a day do not hurt anyone'
(C48: 721)

(62) *Mas se de repente eu conheço que tem um banco de leite materno.*
'But if suddenly I know that there is a milk bank.'
(I9: 287)

Non-referential verbal forms are also the source of constructions that seem to affect the retention of non-expression in BP. This can be seen in two particular constructions, i.e. *é* 'to be-$_{3SG\text{-}PRESENT}$' and *era* 'to be$_{-3SG\text{-}IMPERFECT}$' as can be seen in (63) and (64) respectively.

(63) Inf. - *aí é*
já é
na própria escola é
onde cê tá trabalhando a reciclagem
'Then (it) is
just (it) is
at the school is
where you are working the recycling'
(L5: 284-287)

(64) *Tinha acesso Ideal, **era** nessas festa de formatura que a gente se encontrava.*
'(I) had access to Ideal, (it) was in these graduation parties that we would meet'
(L12: 650)

As illustrated in (63) and (64) above, these constructions do not occur with subjects that cannot be identified due to their irretrievability from the context; rather they appear to be

verbs with no semantic subject, and to have developed into discourse markers or topic markers. For example, the form *é* 'to be$_{\text{-3SG-PRESENT}}$' seems to work as marker of thetic sentences in BP, much like sentences with motion verbs and post-posed subjects as has been argued by Zilles (2000).

It is argued that thetic structures in languages across the world function as presentational structures, that is, they normally "tend to be intransitive, containing predicates indicating the existence or coming into existence of a referent, or the appearance of a referent in the external or internal world of discourse" (Lambrecht 1994, p. 143). It appears to be same case in the data analyzed. These "subjectless" verbs have discourse functions that exceed their functions as predicates of topic-comment clauses. Indeed, it seems that these clauses are behaving as discourse markers, or topic markers in BP. Example (65) illustrates this claim.

(65) *Inf. 1 - AÍ é que,*
é menos indicado,
pra 'tá testando gratuitamente.
'Then (that) is that, (it) is less likely to be testing it for free.'
(C7: 437-439)

In this example, the predicate *é* '(it) is' appears twice and in neither case does it have a subject, and they predicate an evaluative proposition of the forthcoming utterance, i.e. the inflected form appears to display an evaluative stance of what the speaker believes to be true regarding the proposition.

Also, I would like to propose that these forms attain a high level of subjectivity. They reflect the speaker's evaluation of events, places, and attitudes, i.e. what speakers do all the time when conversing with others. Namely, they are personalizing their discourse. In alignment with Langacker's model (1983), the form allows the speaker to displace

themselves from their proposition (see section 4.1.2 for discussion on displacement). They take the perspective of the hearer on the utterance and as such it becomes more subjective. These subjective forms display the evaluations of the speaker towards a proposition as illustrated by (66).

(66) *Inf. 1 - ...o orçamento.*
Inf. 2 - o orçamento de se fazer uma faro=fa,
... com uma salada assim tipo salpicão.
É um jantar.
... É uma farofa junto com o peru.
'Inf. 1 - the budget.
Inf. 2 - The budget to make a farofa,
… with a salad and chicken salad.
(It) is a dinner.
… (it) is a farofa with turkey.'

(Inq. 7: 609-613)

In the first clause the speaker states the nature of the meal to justify her choice of the dish "farofa". Then she goes on to elaborate on what is going to be served for dinner. This construction clearly illustrates the notion of displacement. The speaker wants to say that she thinks the combination is a good choice for dinner (despite the stereotypical conception that you usually do not eat farofa with turkey unless during Christmas), and she does it so by stating that the specific dish is going to be accompanied by another. The speaker makes an evaluation of the proposition without taking part on it. This allows her to invite the hearer to evaluate the proposition from the observer point of view. This is the configuration that Langacker proposes to be ideally subjective and it can be seen in the data by a striking number of tokens of é '(it) is'.

Another function exerted by *é* '(it) is' is that of showing agreement and letting the speaker know that his proposition is understood and/or that he is being followed. In (67) the speaker is talking about a woman who wants to place her child in a boarding school. As the

speaker continues explaining the reasons for such a decision and goes on with her narrative, the hearer maintains interest, thus building up rapport, by using é ‘(it) is’ two times.

(67) Inf. 2 - *não.*
... mas também não se matriculou,
... ela ‘tava querendo botar no semi-internato né?
Mas num disse que ainda ia dar um balanço nas finan=ças,
num sei,
o que,
num sei se va=i ou se num vai.
Inf. 1 - ah é.
ela quer --
ela queria --
Inf. 2 - é
Inf. 1 - botar no semi-internato.,
Inf. 2 - é=
‘Inf. 2 - no.
… but (she) also did not enroll,
… she wanted to send him to a boarding school, right?
But (she) said that she would try to balance her checkbook,
(I) don’t know,
that,
(I) don’t know if it is going to work or not.
Inf. 1 - oh, sure.
She wants --
she wanted –
Inf. 2 - yeah.
Inf. 1 - send him to boarding school.
Inf. 2 - yeah.’

(C7: 589-571)

The remaining constructions do not demonstrate a clear discourse functions, but they do share the characteristics of discourse markers in the making. They do not have any other linguistic elements in their composition, which is consistent with Company Company (2007) in that discourse markers demonstrate these characteristics, which are formal ways of detecting their pragmatic function along with their bleaching of meaning in place for a more discourse meaning.

What must be noted, however, about these constructions as a group is that the ubiquity of non-expression in these high-frequency constructions permeates through the language, creating a robust pattern that becomes ingrained in the minds of speakers and resistant to replacement by another emerging pattern, such as pronominal expression. Such level of fossilization obtained in these constructions is suggestive of a long history of grammaticization of these forms without pronominal subjects to eventually becoming so general that their forms do not show variability of any kind; in fact, these constructions do not take any form of pronominal subjects without sounding ungrammatical or changing their meaning and/or function entirely.

7.6 Discussion

Departing from the premise of construction grammar that 'cognitive representations are affected by the speaker's experience with language' (Bybee & Cacoullos, 2009, p. 187), the constructions illustrated in the last four sections have corroborated such premise in that their high-frequency and their co-occurrence with particular subjects have shaped the way pronominal expression manifests in BP.

One of the major contributions of this work lies in the discussion of the equivalent to 58% of all tokens culled for this study, i.e. the proportion of the data that was discussed in Chapters 4, 6, and 7 (10953/18810). And from these data we have learned that pronominal expression is far more complex than any statistical analysis is able to capture. Each of the high frequency predicates are patterning individually according to their rates or expression, their functions in discourse, and the tenses with which each occurs.

The constructions described in this chapter showed different levels of abstraction, while some of them were fossilized in their function in discourse; others simply appeared to

be highly frequent forms occurring with one or the other type of subject, that is, with pronominal or unexpressed subjects. Although this study has set out to show how these constructions affect pronominal expression in BP, it is not our mission to lay out the function of every single construction revealed here. Rather, I hope to have showed that they indeed affect variant choice based on their co-occurring patterns, and that they should be taken into account in future analyses. In addition, I believe that more work should be done to arrive at appropriate cutoff points that better express what the notion of frequency really is and how it can influence language structure.

8 SUMMARY AND CONCLUSIONS

The present study was designed to determine the effect of a set of factor groups, the effect of the same factor groups when the highest frequency verbs are excluded, and the effect of constructions of subject, verb and tense on the conditioning of pronominal subjects in Brazilian Portuguese. To achieve this goal Variationist theory and Usage-based Linguistics have been employed in this study to examine the pattern of pronominal expression in 1sg, 2sg, and 3sg subjects in Brazilian Portuguese. The Variationist comparitve method was employed to test the differences between the three persons and these datasets without the most highly frequent verbs.

One of the most significant findings to emerge from this study was that there are factor groups that affect each person in different ways, that is, , the hierarchy of constraints within a factor group varies according to the person being examined. A case in point is the factor group **TAM** which was selected as significant in both 2sg and 3sg analyses. In both analyses the **IMPERFECT** favors pronominal expressions, but the **PRETERIT** and the **PRESENT** have differing patterning. The former favors expression with 2sg subjects and disfavors it with 3sg subjects, while the latter favors expression with 3sg subjects and disfavors it with 2sg subjects. Such changes in the ordering of factors within a factor group are indicative of the differences between the three subjects in their patterning with pronominal subjects in BP, and of the kind of information that is lost when different persons are considered together.

Another finding obtained from the comparison of statistical analyses on the three persons of speech is observed in that not all factor groups condition the three persons in the same way, that is, each person shows its own group of factors and within that, its own direction of effect, that affect the patterning of subject expression. This is evidenced in the

fact that for both 1sg and 2sg subjects **VERB CLASS** plays a strong role in variant choice, while for 3sg subjects this factor group is not even selected as significant. In addition, even those factor groups that do affect more than one person do not do so in the same way, that is In short, the comparison between the three persons reveals only one factor group that conditions pronominal expression across the three persons, i.e., **CLAUSE TYPE**, and the other factor groups are divided in their significance across the three persons. For instance, **VERB CLASS** and **MORPHOLOGICAL IRREGULARITY** are only significant in 1sg and 2sg subjects; **TAM** and **MODAL** are significant in 2sg and 3sg subjects; **DISCOURSE CONTINUITY** and **POLARITY** are significant in 1sg and 3sg subjects. Such discreet distribution among these factor groups according to person is robust support to the claim that these subjects need to be analyzed separately.

A major contribution of this study can be observed in the findings obtained from the examination of the effect of high-frequency predicates. Firstly, we observe that only a handful of predicates, 25 distributed across the three persons (18 verbs with 1sg, 21 verbs with 2sg, and 16 verbs with 3sg) account for the cutoff of one percent of the dataset to be considered of high frequency. Furthermore, these 25 verbs account for 66% of the data for 1sg, 73% of the data for 2sg, and 56% of the data for 3sg subjects, which in itself is very suggestive of their effect on the syntactic pattern being examined here. Statistical analyses were conducted on the portion of the data remaining for each person separately to identify whether there were any differences in the factor groups that condition the realization of pronominal subjects in these datasets. The findings showed a pronounced difference between these new analyses and the ones with all predicates included. Firstly, it was noted that there was a considerable increase in the rates of expression for 1sg and 2sg subjects, suggesting

that pronominal expression is more entrenched with these subjects. Secondly, variation in the significance of factor groups is observed in that factor groups that were selected as significant in earlier analyses are not selected as significant in the analysis without the most frequent predicates; other factor groups showed a weak effect in other analyses and they became stronger in this analyses; or other factor groups that were strong predictors in earlier analyses are no longer significant predictors of pronominal expression. These findings suggest that different portions of the data are constrained by different forces in their distribution of pronominal subjects. Lastly, within factor groups selected as significant in both analyses with and without the most frequent predicates, there is a change in their ordering of factors, reinforcing the distinct behavior of the two data portions. In sum, each person is behaving markedly from the others at several levels of analyses, and these differences pose an important methodological question on how to examine them.

A third finding obtained from these analyses has its roots in the construction identified and analyzed in chapter 7. The constructions that categorically occurred without pronominal subjects and the high-frequency predicates amount to 58% of all clauses originally extracted (10953/18810). These constructions showed different levels of abstraction from highly formulaic forms to frequent forms that did not exhibit any formulaicity but formed frequent collocations in the data. These patterns constitute combination of subject (person + expression or lack thereof), predicate and tense, mostly present, with the exception of *era* 'to be$_{\text{-3SG-IMPERFECT}}$' and *entendeu* '(you) understand$_{\text{-PRETERIT}}$'. The remaining constructions occurred with the present tense, e.g. *sei* '(I) know$_{\text{-PRESENT}}$', *entende* '(you) understand$_{\text{-PRESENT}}$', and *tem* 'have$_{\text{-3SG-PRESENT}}$' to name a few. While the constructions emerging from 1sg, 2sg, and 3sg collocations with predicates and tenses can be

considered as part of the structure SUBJECT + PREDICATE + TENSE, the constructions that categorically occur without pronominal subjects may not. This is because these constructions have grammaticized over time and their constituent structure no longer factors in the SUBJECT part of the structure.

The findings obtained thus far have shown that, in spontaneous discourse, variant choice is affected to a large degree by patterns of frequently co-occurring subjects, verbs and tenses, and that such examination should be a common practice within linguistic analysis. More importantly, the frequently co-occurring constructions of subjects, verbs and tenses, especially those that fall out of the envelope of variation, should be carefully examined because their high frequency may have a larger impact on variant choice than statistically significant factors.

In addition to proposing a further avenue for investigating subject expression in Brazilian Portuguese, and to a greater extent, Romance languages, this work highlights the role of frequency in language change as a factor that must be examined and accounted for. In short, this dissertation shows that by observing different ranges of frequency, one can observe different linguistic patterning of structures, and how changes spread across languages. Such observations are a major contribution for a more holistic understanding of language change as it progresses.

REFERENCES

Armstrong, D. F., Stokoe, W. C., & Wilcox, S. E. (1995). *Gesture and the nature of language*. Cambridge: Cambridge University Press.

Ashby, W. J., & Bentivoglio, P. (1993). Preferred argument structure in spoken French and Spanish. *Language Variation and Change, 5*, 61-76.

Ávila-Shah, B. I. (2000). Discourse connectedness in Caribbean Spanish. In A. Roca (Ed.), *Research on Spanish in the United States* (pp. 238-251). Somerville: Cascadilla Press.

Bailey, G. (2002). Real and apparent time. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 312-332). Oxford: Blackwell Publishers.

Barbosa, P. (1995). *Null Subjects.* Ph.D. Ph.D. Dissertation, MIT.

Barbosa, P., Duarte, M. E. L., & Kato, M. A. (2005). Null Subjects in European and Brazilian Portuguese. *Journal of Portuguese Linguistics, 4*, 11-52.

Barddal, J. (2006). Construction-specific properties of syntactic subjects in Icelandic and German. *Cognitive Linguistics, 17*(1), 39-106.

Barddal, J., & Eythórsson, T. (2003). The change that never happened: the story of oblique subjects. *Journal of Linguistics, 39*, 439-472.

Barddal, J., Kristoffersen, K. E., & Sveen, A. (2011). West Scandinavian ditransitives as a family of constructions: With special attention to the Norwegian 'V-REFL-NP' construction. *Linguistics, 49*(1), 53-104.

Barlow, M., & Kemmer, S. (2000). *Usage-based models of language*. Stanford, CA: CSLI Publications.

Barrenechea, A. M., & Alonso, A. (1977). Los pronombres personales sujetos en el español hablado en Buenos Aires. In J. M. Lope Blanch (Ed.), *Estudios sobre el español hablado en las ciudades principales de América* (pp. 333-349). México: Universidad Nacional Autónoma de México.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*(1), 92-111.

Bentivoglio, P. (1983). Topic continuity and discontinuity: A study of Latin-American spoken Spanish. In T. Givón (Ed.), *Topic continuity in discourse: A quantitative cross-language study* (pp. 255-312). Amsterdam/Philadelphia: John Benjamins.

Bentivoglio, P. (1987). *Los sujetos pronominales de primera persona en el habla de Caracas*. Caracas: Universidad Central de Venezuela.

Bergan, B. K., & Chang, N. (2005). Embodied Construction Grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction grammars: cognitive grounding and theoretical extensions* (pp. 147-190). Amsterdam/Philadelphia: John Benjamins.

Biber, D. (1995). *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Philadelphia: John Benjamins.

Bybee, J. (1994). The grammaticization of zero: Asymmetries in tense and aspect systems. In W. Pagliuca (Ed.), *Perspectives on grammaticalization* (pp. 235-254). Amsterdam: John Benjamins.
Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*(5), 425-455.
Bybee, J. (2001). *Phonology and Language Use*: Cambridge University Press.
Bybee, J. (2002a). Main clauses are innovative, subordinate clauses are conservative: Consequences for the nature of constructions. In J. Bybee & M. Noonan (Eds.), *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson* (pp. 1-17). Amsterdam / Philadelphia: John Benjamins.
Bybee, J. (2002b). Sequentiality as the Basis of Constituent Structure. In T. Givón & B. F. Malle (Eds.), *The evolution of langauge out of pre-language* (pp. 109-134). Amsterdam: John Benjamins.
Bybee, J. (2002c). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change, 14*(3), 261-290.
Bybee, J. (2003). Mechanisms of change in grammaticalization: The role of frequency. In B. D. Joseph & R. D. Janda (Eds.), *Handbook of historical linguistics* (pp. 602-623). Oxford: Blackwell Publishers.
Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711-733.
Bybee, J. (2010). *Language, Usage and Cognition*: Cambridge University Press.
Bybee, J., & Cacoullos, R. T. (2009). The role of prefabs in grammaticization: How the particular and the general interact in language change. In R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (Eds.), *Formulaic Language* (Vol. 1. Distribution and historical change, pp. 187-217). Amsterdam: John Benjamins.
Bybee, J., & Dahl, Ö. (1989). The creation of tense and aspect systems in the langauges of the world. *Studies in Language, 13*(1), 51-103.
Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The Evolution of Grammar: Tense, aspect, and Modality in the Languages of the World*. Chicago and London: The University of Chicago Press.
Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics, 37*(4), 575-596.
Bybee, J., & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language, 58*, 265-289.
Bybee, J., & Thompson, S. A. (1997). Three frequency effects in syntax. *Berkeley Linguistics Society, 23*, 378-388.
Cameron, R. (1992). *Pronominal and null subject variation in Spanish: Constraints, dialects, and functional compensation.* PhD, University of Pennsylvania, Philadelphia, PA.
Cameron, R. (1994). Switch reference, verb class and priming in a variable syntax. *Papers from the Regional Meeting of the Chicago Linguistic Society: Parasession on variation in linguistic theory, 30*(2), 27-45.
Cameron, R. (1995). The scope and limits of switch reference as a constraint on pronominal subject expression. *Hispanic Linguistics, 6/7*, 1-27.
Cameron, R. (1996). Accessibility theory and specificity of reference in Spanish. *Papers from the Regional Meeting of the Chicago Linguistic Society, 32*(2), 13-27.

Cameron, R., & Flores-Ferrán, N. (2003). Perseveration of subject expression across regional dialects of Spanish. *Spanish in Context, 1*(1), 41-65.
Cameron, R., & Flores-Ferrán, N. (2004). Preservation of subject expression across regional dialects of Spanish. *Spanish in Context, 1*(1), 41-65.
Castilho, A. d. T. (1987). A elipse do sujeito no português culto falado em São Paulo. *Estudos Lingüísticos, 14*, 32-40.
Cavalcante, M. A. (2001). *O sujeito pronominal em Alagoas e no Rio de Janeiro: um caso de mudança em progresso.* Ph.D. Ph.D. Dissertation, UFAL, Maceió.
Cedergren, H., & Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language, 50*, 333-355.
Chafe, W. (1994). *Discourse, consciousness and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.
Chomsky, N. (1965). *Aspects of .the Theory on Syntax*. Cambridge: MIT Press.
Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding.* MA.
Chomsky, N. (1991). Some notes on economy of derivation and representation. In R. Freidin (Ed.), *Principles and Parameters in Comparative Grammar* (pp. 417-454). Cambridge, MA: MIT Press.
Christy, C. (1983). *Uniformitarianism in Linguistics*. Amsterdam and Philadelphia: John Benjamins.
Clements, J. C. (2006). Null direct objects in Spanish. In J. C. Clements & J. Yoon (Eds.), *Functional approaches to Spanish syntax: Lexical semantics, discourse and transitivity* (pp. 203-226). New York: Palgrave MacMillan.
Company Company, C. (2007). Subjectification of verbs into discourse markers: Semantic-pragmatic change only? In N. Delbecque & B. Cornillie (Eds.), *Modalization and pragmaticalization*. Amsterdam: John Benjamins.
Comrie, B. (1976). *Aspect*. Cambridge: Cambridge University Press.
Comrie, B. (1981). *Language Universals and Language Typology*: University of Chicago Press.
Coulmas, F. (2005). *The study of Speaker's Choices*. Cambridge: Cambridge University Press.
Croft, W. (2001). *Radical Construction Grammar*. Oxford: Oxford University Press.
Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambrisge University Press.
Cunha, C., & Cintra, L. (1985). *Gramática do Português Contemporâneo*. Lisboa: Edições João de Sá da Costa.
Cunha, S. (2003). *Terras de preto no Maranhão.* Ph.D. Ph.D. Dissertation, USP, São Paulo, SP.
Deutscher, G. (2000). *Syntactic Change in Akkadian: The Evolution of Sentential Complementation*: Oxford University Press.
Dixon, R. M. W. (1979). Ergativity. *Language, 55*(1), 59-138.
Dixon, R. M. W. (2009). *Basic Linguistic Theory Volume 2: Grammatical Topics*: Oxford University Press.
Du Bois, J. W. (1987). The discourse basis of ergativity. *Language, 63*, 805-855.

Du Bois, J. W. (2003). Discourse and grammar. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (Vol. 2, pp. 47-88). Mahwah, NJ: Lawrence Erlbaum Associates.

Du Bois, J. W., Kumpf, L. E., & Ashby, W. J. (Eds.). (2003). *Preferred Argument Structure: Grammar as architecture for function*. Amsterdam / Philadelphia: John Benjamins.

Duarte, M. E. L. (1993). Do Pronome Nulo ao Pronome Pleno: A Trajetória do Sujeito no Português do Brasil. In I. Roberts & M. Kato (Eds.), *Português Brasileiro: uma viagem diacrônica* (pp. 107-128). São Paulo: Pontes.

Duarte, M. E. L. (2000). The Loss of the 'Avoid Pronoun' Principle in Brazilian Portuguese. In M. A. Kato & E. V. Negrão (Eds.), *Brazilian Portuguese and the null subject parameter* (pp. 17-36). Madrid: Iberoamericana.

Duarte, M. E. L. (2003). A Evolução na Representação do Sujeito Pronominal em Dois Tempos. In M. C. d. Paiva & M. E. L. Duarte (Eds.), *Mudança Lingüística em tempo Real* (pp. 115-128). Rio de Janeiro: FAPERJ.

Dutra, R. (1987). The hybrid S-category in Brazilian Portuguese: Some implications for word order. *Studies in Language, 11*(1), 163-180.

Enríquez, E. V. (1984). *El pronombre personal sujeto en la lengua española hablada en Madrid*. Madrid: Consejo Superior de Investigaciones Científicas, Instituto Miguel de Cervantes.

Enríquez, E. V. (1986). La presencia de los pronombres personales sujeto en el mundo hispánico: Estudio comparativo. *Anuario de Letras, 24*, 48-70.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20*(1), 29-62.

Faraco, C. A. (1996). O tratamento *você* em português: uma abordagem histórica. *Fragmenta, 13*, 51-82.

Ferreira, M. B. (2000). *Argumentos Nulos em Português Brasileiro.* MA MA Thesis, Universidade Estadual de Campinas, Campinas, SP.

Fillmore, C. J. (1975). Some problems for case grammar. *Langages, 9*(38), 65-80.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language, 64*(3), 501-538.

Flores-Ferrán, N. (2002). *Subject personal pronouns in Spanish narratives of Puerto Ricans in New York City: A sociolinguistic perspective*. Munich: Lincom Europa.

Galves, C. (2000). Agreement, Predication, and Pronouns in the History of Portuguese. In J. Costa (Ed.), *Portuguese Syntax* (pp. 143-168). New York: Oxford.

Givón, T. (1971). Historical syntax and synchronic morphology: An archeologist's field trip. *Chicago Linguistics Society, 7*, 394-415.

Givón, T. (1976). Topic, pronoun, and grammatical agreement. In C. Li (Ed.), *Subjects*.

Givón, T. (1983a). Topic continuity and word-order pragmatics in Ute. In T. Givón (Ed.), *Topic continuity in discourse: A quantitative cross-language study* (pp. 141-214). Amsterdam/Philadelphia: John Benjamins.

Givón, T. (1983b). Topic continuity in discourse: An introduction. In T. Givón (Ed.), *Topic continuity in discourse* (pp. 3-41). Amsterdam: John Benjamins.

Givón, T. (1984). The Pragmatics of referentiality. *Georgetown University Round Table on Languages and Linguistics*, 120-138.

Givón, T. (1987). Beyond foreground and background. In R. S. Tomlin (Ed.), *Coherence in grounding and discourse* (pp. 175-188). Amsterdam / Philadelphia: John Benjamins.

Givón, T. (1995). *Functionalism and grammar*. Amsterdam: John Benjamins.
Givón, T. (2001). *Syntax: An Introduction* (Vol. I). Amsterdam/Philadelphia: John Benjamins.
Givón, T. (Ed.). (1983c). *Topic continuity in discourse*. Amsterdam: John Benjamins.
Goddard, C. (2005). *The Languages of East and Southeast Asia: An Introduction*: Oxford University Press.
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics, 15*(3), 289-316.
Guy, G. (1988). Advanced VARBRUL analysis. In K. Ferrara, B. L. Brown & J. Walters (Eds.), *Linguistic change and contact: Proceedings of the annual conference on New Ways of Analyzing Variation* (Vol. 16, pp. 124-136).
Guy, G. R. (1993). The quantitative analysis of linguistic variation. In D. R. Preston (Ed.), *American dialect research* (pp. 223-249). Amsterdam: John Benjamins.
Haiman, J. (1994). Ritualization and the Development of Language. In W. Pagliuca (Ed.), *Perspectives on Grammaticalization* (pp. 3-28). Amsterdam/Philadelphia: John Benjamins.
Halliday, M. A. K. (1994). *An introduction to functional grammar*. London: Edward Arnold.
Halliday, M. A. K. (2004). *An Introduction to Functional Grammar* (3rd ed.). New York, NY: Oxford University Press.
Harbert, W. (2007). *The Germanic Languages*. Cambridge: Cambridge University Press.
Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, 1041-1070.
Hay, J., & Baayen, H. (2002). Parsing and productivity. *Yearbook of morphology, 2001*, 203-235.
Hay, J., & Baayen, H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences, 9*(7), 342-348.
Heine, B., & Kuteva, T. (2007). *The Genesis of Grammar: A Reconstruction*. New York: Oxford.
Hochberg, J. G. (1986). Functional compensation for /s/ deletion in Puerto Rican Spanish. *Language, 62*, 609-621.
Hopper, P. J. (1979). Aspect and foregrounding in discourse. In T. Givón (Ed.), *Discourse and syntax* (pp. 213-241). New York: Academic Press.
Hopper, P. J. (1987). Emergent grammar. *Proceedings of the Berkeley Linguistics Society, 13*, 139-157.
Hopper, P. J. (1988). Emergent grammar and the a priori grammar postulate. In D. Tannen (Ed.), *Linguistics in context* (pp. 117-134). Norwood, NJ: Ablex.
Hopper, P. J. (1998). Emergent Grammar. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (pp. 155-175). Mahwah, NJ: Lawrence Erlbaum.
Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization* (2nd ed.). Cambridge: Cambridge University Press.

Huang, C. T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic Inquiry, 15*(4), 531-574.
Ilari, R., Franchi, C., Moura Neves, M. H. d., & Possenti, S. (1996). Os pronomes pessoais do Português falado: Roteiro para a análise. In A. T. d. Castilho & M. Basilio (Eds.), *Gramática do Português falado: Estudos descritivos* (Vol. 4). São Paulo, Brasil: Editora de Unicamp.
Jackendoff, R. (2004). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York, NY: Oxford Uiversity Press.
Jaeggli, O., & Safir, K. J. (1989). The Null Subject Parameter and Parametric Theory. In O. Jaeggli & K. J. Safir (Eds.), *The Null Subject Parameter* (Vol. 15, pp. 1-44). Dordrecht: Kluwer Academic Publishers.
Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*(2), 137-194.
Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of Neural Science*. New York: McGraw-Hill.
Kato, M. A. (1996). Padrões de predição no português falado no Brasil. In M. A. Kato (Ed.), *Gramática do português falado* (Vol. V, pp. 201-274). Campinas: Fapesp/Editora da Unicamp.
Kato, M. A. (1999). Strong and weak pronominals in the null subject parameter. *Probus, 11*(1), 1-37.
Kato, M. A. (2000). The Partial Pro-drop Nature and the restricted VS Order in Brazilian Portuguese. In M. A. Kato & E. V. Negrão (Eds.), *Brazilian Portuguese and the null subject parameter* (pp. 223-254). Madrid: Iberoamericana.
Kay, P. (1978). Variable rules, community grammar, and linguistic change. In D. Sankoff (Ed.), *Linguistic variation: models and methods*. New York: Academic Press.
Kemmer, S., & Barlow, M. (2000). Introduction: A Usage-Based Conception of Language. In M. Barlow & S. Kemmer (Eds.), *Usage-Based Models of Language* (pp. vii-xxviii). Stanford, California: CSLI.
Kenstowicz, M. (1989). The Null Subject Parameter in Modern Arabic Dialects. In O. Jaeggli & K. J. Safir (Eds.), *The Null Subject Parameter* (Vol. 15, pp. 263-276). Dordrecht: Kluwer Academic Publishers.
Kövecses, Z., & Szabo, P. (1996). Idioms: A view from cognitive semantics. *Applied Linguistics, 17*(3), 326-355.
Krug, M. (1998). String frequency: A cognitive motivating factor in coalescence, language processing and linguistic change. *Journal of English Linguistics, 26*, 286-320.
Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language, 45*(4), 715-762.
Labov, W. (1971). Some principles of linguistic methodology. *Language in Society, 1*(1), 97-120.
Labov, W. (1972a). *Sociolinguistic patterns*. Oxford: Basil Blackwell.
Labov, W. (1972b). Some principles of linguistic methodology. *Language in Society, 1*(1), 97-120.
Labov, W. (1994). *Principles of Linguistic Change, Volume I: Internal Factors*. Oxford: Blackwell.
Labov, W. (2001). *Principles of Linguistic Change: Social Factors* (Vol. 2). Oxford: Blackwell Publishers.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.
Lambrecht, K. (2001). A framework for the analysis of cleft constructions. *Linguistics, 39*(3), 463-516.
Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford: Stanford University Press.
Langacker, R. W. (2000). A Dynamic Usage-Based Model. In M. Barlow & S. Kemmer (Eds.), *Usage-Based Models of Language* (pp. 1-63). Stanford, California: CSLI.
Levinson, S. C. (1987). Pragmatics and the grammar of anaphora: A partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics, 23*(2), 379-434.
Li, C. N., & Thompson, S. A. (1976). Subject and topic: A new typology of language. In C. N. Li (Ed.), *Subject and topic* (pp. 459-489). New York: Academic Press.
Lira, S. d. A. (1982). *Nominal, Pronominal and Zero Subject in Brazilian Portuguese.* Ph.D. Dissertation, University of Pennsylvania.
MacWhinney, B. (2005). A Unified Model of Language Acquisition. In J. F. Kroll & A. M. B. Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 49-67). Oxford: Oxford University Press.
McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6*(11), 465-472.
Milroy, J. (1992). *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford/Cambridge: Blackwell.
Modesto, A. T. T. (2006). *Formas de tratamento no português brasileiro: A Alternância tu/você na cidade de Santos-SP.* MA Thesis, Universidade de São Paulo, São Paulo.
Modesto, M. (2000a). Null Subjects Without 'Rich' Agreement. In M. A. Kato & E. V. Negrão (Eds.), *Brazilian Portuguese and the null subject parameter* (pp. 147-174). Madrid: Iberoamericana

Modesto, M. (2000b). *On the Identification of Null Arguments.* Ph.D. Ph.D. Dissertation, University of Southern California.
Monteiro, J. L. (1990). Variação no uso dos pronomes pessoais no portugues do Brasil. *Verba, 17*, 145-157.
Monteiro, J. L. (1994a). PORCUFORT (Corpus of Educated Oral Portuguese from Fortaleza), from http://www.geocities.com/Paris/Cathedral/1036/
Monteiro, J. L. (1994b). *Pronomes pessoais do Português Brasileiro*. Fortaleza, CE: Editora Vozes.
Morales, A. (1997). La hipótesis funcional y la aparición de sujeto no nominal: El español de Puerto Rico. *Hispania, 80*(1), 153-165.
Naro, A. J. (1981). The social and structural dimensions of a syntactic change. *Language, 57*(1), 63-98.
Naro, A. J. (1992). Modelos quantitativos e tratamento estatístico. In M. C. Mollica (Ed.), *Introdução à sociolinguística variacionista* (pp. 18-25). Rio de Janeiro: Cadernos Didáticos.
Naro, A. J., & Scherre, M. M. P. (1991). Linguistic variation and change: Change and counterchange in a speech community. *Cadernos de Estudos Linguisticos, 20*, 9-16.
Negrão, E. V., & Müller, A. L. (1996). As Mudanças no Sistema Pronominal do Português Brasileiro: Substituição ou Especialização de Formas? *D.E.L.T.A, 12*(1), 125-152.

Negrão, E. V., & Viotti, E. (2000). Brazilian Portuguese as a Discourse-Oriented Language. In M. A. Kato & E. V. Negrão (Eds.), *Brazilian Portuguese and the Null Subject Parameter* (pp. 105-126). Madrid: Iberoamericana.

Newmeyer, F. J. (1998). *Language, form and function*. Cambridge: MIT Press.

Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language, 79*(4), 682-707.

Nunberg, G., Sag, I. A., & Wason, T. (1994). Idioms. *Language, 70*(3), 491-538.

Ono, T., & Thompson, S. A. (1997). Deconstructing 'zero anaphora' in Japanese. *Berkeley Linguistics Society, 23*, 481-491.

Otheguy, R., Zentella, A. C., & Livert, D. (2007). Language and Dialect Contact in Spanish in New York: Toward the Formation of a Speech Community. *Language, 83*(4), 770-802.

Paredes Silva, V. L. (1993). Subject omission and functional compensation: Evidence from written Brazilian Portuguese. *Language Variation and Change, 5*(1), 35-49.

Paredes Silva, V. L. (2003). Motivações Funcionais no Uso do Sujeito Pronominal: Uma Análise em Tempo Real. In M. d. C. Paiva & M. E. L. Duarte (Eds.), *Mudança Lingüística em Tempo Real* (pp. 97-114). Rio de Janeiro/RJ: Contra Capa.

Perini, M. A. (2002). *Modern Portuguese: A Reference Grammar*. New Haven and London: Yale University Press.

Pickering, M. J., Branigan, H. P., Cleland, A. A., & Stewart, A. J. (2000). Activation of Syntactic Information During Language Production. *Journal of Psycholinguistic Research, 29*(2), 205-216.

Platzack, C. (1987). The Scandinavian Languages and the Null-Subject Parameter. *Natural Language and Linguistic Theory, 5*(3), 377-401.

Poplack, S. (1993). Variation theory and language contact: Concepts, methods and data. In D. R. Preston (Ed.), *American dialect research* (pp. 251-286). Amsterdam: John Benjamins.

Poplack, S. (2001). Variability, frequency, and productivity in the irrealis domain in French. In J. Bybee & P. J. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 405-428). Amsterdam / Philadelphia: John Benjamins.

Poplack, S., & Tagliamonte, S. (2001). *African American English in the diaspora*. Massachusetts: Blackwell Publishers.

Raposo, E. P. (1998). Some Observations on the Pronominal System of Portuguese. *CatWPL, 6*, 59-93.

Rizzi, L. (1986). Null Objects in Italian and the Theory of pro. *Linguistic Inquiry, 17*(3), 501-557.

Rollemberg, V., Andrade, C., Lopes, C., & Matos, C. (1991). Os Pronomes Pessoais e a Indeterminação do Sujeito na Norma Culta de Salvador. *Estudos, 11*, 53-74.

Rousseau, P., & Sankoff, D. (1978). Advances in variable rule methodology. In D. Sankoff (Ed.), *Linguistic variation: models and methods* (pp. 57-68). New York: Academic Press.

Sankoff, D. (1988a). Sociolinguistics and syntactic variation. In F. Newmeyer (Ed.), *Linguistics: The Cambridge survey (Vol. 4, Language: The socio-cultural context)* (pp. 140-161). Cambridge: Cambridge University Press.

Sankoff, D. (1988b). Variable rules. In U. Ammon, N. Dittmar & K. J. Mattheier (Eds.), *Sociolinguistics: An international handbook of the science of language and society* (Vol. 2, pp. 984-997). Berlin / New York: Walter de Gruyter.

Sankoff, D., & Labov, W. (1979). On the uses of variable rules. *Language in Society, 8*(2), 189-222.
Sankoff, D., Tagliamonte, S., & Smith, E. (2005). GOLDVARB X: A multivariate analysis application for Macintosh and Windows, from http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm
Sapir, E. (1949). *Language: an introduction to the study of speech*. New York: Harcourt, Brace & Co.
Scheibman, J. (2001). Local Patterns of Subjectitvity in Person and Verb Type in American English Conversation. In J. Bybee & P. Hopper (Eds.), *Frequency and the emmergence of linguistic structure* (pp. 61-89). Amsterdan/Philadelphia: John Benjamins.
Silva-Corvalán, C. (1982). Subject expression and placement in Mexican-American Spanish. In J. Amastae & L. Elías Olivares (Eds.), *Spanish in the United States: Sociolinguistic aspects* (pp. 93-120). New York: Cambridge University Press.
Silva-Corvalán, C. (1994). *Language contact and change: Spanish in Los Angeles*. Oxford: Claredon Press.
Silva-Corvalán, C. (1997). Avances en el estudio de la variación sintáctica: La expresión del sujeto. *Cuaderno del Sur: Letras, Homenaje a Beatriz Fontanella de Weinberg, 27*, 35-49.
Silva-Corvalán, C. (2001). *Sociolingüística y pragmática del español*. Washington DC: Georgetown University Press.
Silva, V. L. P., Santos, G. M. d., & Ribeiro, T. d. O. (2000). Variação na 2a pessoa: o pronome sujeito e a forma do imperativo. *Niterói, 9*, 115-123.
Silva, V. L. P. d. (1996). Quando escrita e fala se aproximam: Pronomes de terceira pessoa em cartas pessoais. In A. T. d. Macedo, C. Rocanrati & M. C. Mollica (Eds.), *Variação e Discurso* (pp. 85-96). Rio de janeiro: Tempo Brasileiro.
Silveira, A. S. (2007). Patterns of Subjectivity in Spoken Brazilian Portuguese. In C. L. Pinheiro & K. M. R. Campelo (Eds.), *Português Oral Culto de Fortaleza: estudos descritivos* (pp. 200-247). Fortaleza, CE: Eduece.
Silveira, A. S. (2008). Frequency Effects, Specialization of Forms, and Subject Expression in Brazilian Portuguese. *Proceedings of the High Desert Linguistics Society VII Linguistics Conference*, 1-19.
Sinclair, J. M. (1991). *Corpus, concordance and collocation*. Oxford: Oxford University Press.
Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
Tarallo, F. (1993). Diagnosticando uma gramática brasileira: o português d'aquem e d'além mar no final do século XX. In I. Roberts & M. Kato (Eds.), *Português brasileiro: uma viagem diacrônica*. Campinas, SP: UNICAMP Editora.
Thompson, S. A. (2002). 'Object Complements' and conversation: Towards a realistic account. *Studies in Language, 26*(1), 125-163.
Torres Cacoullos, R. (2006). Relative frequency in the grammaticization of collocations: Nominal to concessive *a pesar de*. In T. L. Face & C. A. Klee (Eds.), *Selected proceedings of the 8th Hispanic Linguistics Symposium* (pp. 37-49). Somerville, MA: Cascadilla Proceedings Project.
Torres Cacoullos, R., & Aaron, J. E. (2003). Bare English-origin nouns in Spanish: Rates, constraints and discourse functions. *Language Variation and Change, 15*(3), 289-328.

Torres Cacoullos, R., & Travis, C. E. (2010). Variable *Yo* Expression in New Mexico: English Influence? In S. Rivera-Mills & D. J. Villa (Eds.), *Spanish of U.S. Southwest: A language in transition* (pp. 185-206). Madrid: Iberoamericana.

Torres Cacoullos, R., & Walker, J. A. (2009). On the Persistence of Grammar in Discourse Formulas: A Variationist Study of *that*. *Linguistics, 47*, 1-43.

Travis, C. E. (2005). The *yo-yo* effect: Priming in subject expression in Colombian Spanish. In R. Gess & E. J. Rubin (Eds.), *Selected papers from the 34th Linguistic Symposium on Romance Languages (LSRL), Salt Lake City, 2004* (pp. 329-349). Amsterdam/Philadelphia: John Benjamins.

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., & Silveira, A. S. (2009). The role of frequency in first person plural variation in Brazilian Portuguese. *Studies in Hispanic and Lusophone Linguistics, 2*(2), 347-376.

Turan, Ü. D. (1995). *Null vs. Overt subjects in Turkish Discourse: A Centering Analysis.* Ph.D. Ph.D. Dissertation, University of Pennsylvania.

Van Lancker, D., Kreiman, J., & Bolinger, D. (1988). Anticipatory lengthening. *Journal of Phonetics, 16*(3), 339-347.

Weinreich, U., Labov, W., & Herzog, M. I. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics: A symposium* (pp. 95-188). Austin, TX: University of Texas Press.

Wolfram, W. (1993). Identifying and interpreting variables. In D. R. Preston (Ed.), *American dialect research* (pp. 193-221). Amsterdam: John Benjamins.

Wray, A. (2000). Formulaic Sequences in Second Language Teaching: Principle and Practice. *Applied Linguistics, 21*(4), 463-489.

Zilles, A. M. S. (2005). The development of a new pronoun: The linguistic and social embedding of *a gente* in Brazilian Portuguese. *Language Variation and Change, 17*, 19-53.