SYNTACTIC VARIATION: THE CASE OF COPULA CHOICE IN LIMÓN, COSTA RICA

Jorge Aguilar-Sánchez

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of TESOL and Applied Linguistics and
the Department of Spanish and Portuguese of the College of Arts and Sciences,
Indiana University

November 2009
Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Manuel Díaz-Campos, Ph.D.

Kimberly Geeslin, Ph.D.

Harry L. Gradman, Ph.D.

Beverly Hartford, Ph.D.

Ken Kelley, Ph.D.

September 18, 2009
Les dedico esta disertación a mis padres y a toda mi familia por todo el apoyo y amor incondicional brindado durante estos nueve años and to my partner Dan and my other family, the McNeelys, for their support and unconditional love during this process. ¡Los amo a todos! I love you all!

Llelli y Papá por fin se les cumplió la predicción, ahora tenemos doctores, masters, licienciados, ingenieros y demás en nuestra familia.

A mis abuelas Pacífica y Ninfa que no me pudieron ver alcanzar este sueño en vida, pero que sé que siempre conté con su apoyo. Las amo mucho!

Jorge
Acknowledgements

At this time I would like to thank all the people who helped me during the stages of this dissertation. First of all, I would like to thank Ivannia, la prima, for all her help and support during the data collection stages and beyond. Cito and Ige for sharing their home with me to be able to work on this project and Doña Mireya, Gino, and Soley as well for allowing me to share part of my life with them. The many members of my nuclear family for putting up with my stress moods while visiting them, I love you all. I would like to thank Eli Winkler for her support over the years and for making part of the data collection for this dissertation possible. Special thanks go to the people of Limón for being so close to my heart and for always being willing to help me with my research projects. I will always be grateful to the committee members for all their insightful comments and guidance, without whom I would have not been able to accomplish my goals and to the Department of Spanish and Portuguese for the AIship given to me over the years; to the staff in the Department of Spanish and Portuguese and in TESOL and Applied Linguistics for making all administrative processes less stressful; to Fulbright-LASPAU for making the first steps in this process possible; to the friends I left in Costa Rica whose friendship is still solid despite the distance, gracias Adrian, Lucy, Doña Mari y Juanca, Chalo, Juanca Gutierrez, Myrna, Bianchi, e Inés; to my friends in Bloomington, Paul, Moraima, Virginia, Sandy, Stephen, Zack, Sarah, Erin, Justine, Ben, Marda, Yudis, Rodolfo, Renato and colleagues who helped in one way or another. Thank you!

I am especially thankful to my professors Manuel and Kim for pushing me to be always better and to follow sound research design; to Ken for introducing me to the field of statistics in a way that I would not have imagined I could explore, it really made a difference; and to Harry and Beverly, professors Gradman and Hartford, for believing that I could become a PhD and giving me the opportunity to further my studies in TESOL and Applied Linguistics.
Finally, most of my gratitude goes to Dan, my partner, for keeping me sane and putting up with me during these nine years. I would not have made it without your support and love. I love you!
Various studies of copula (i.e., *ser* and *estar*) choice in Spanish in contact situations among Spanish-English bilinguals, in contexts where Spanish is not an official language (i.e., Southwestern United States) have shown that copula selection is undergoing change. The change from traditional usage is that the copula *estar* is becoming more accepted in an extended pre-adjectival context, especially with adjectives of size, physical appearance, age, and description and evaluation: this change was initially attributed to contact with English (Silva-Corvalán, 1986). This claim has been extended to monolingual Spanish by Gutiérrez (1992, 1994) because he found that monolingual Mexicans from the same social class as those studied by Silva-Corvalán (1986) showed the same behavior with regard to copula choice. Furthermore, Díaz-Campos & Geeslin (in press) found evidence of this phenomenon in an analysis of the spoken Spanish of Caracas, a context where Spanish is the official language and has no contact with English. Previous studies on the Spanish copula have mostly investigated settings where either Spanish is the only language spoken and the variety has been categorized as monolingual (e.g., Caracas, Venezuela), where Spanish is not the official language (e.g., Los Angeles), and contexts where Spanish is spoken widely, but English has an important role in formal education (e.g., Puerto Rico). Another context in which copula choice in Spanish has been investigated is the acquisition of Spanish as a foreign or second language. These studies have been conducted in settings where Spanish is not an official language (e.g., United States). There is a lack of evidence related to copula choice in contexts in which Spanish is an official language and English is not. The premise of this work is to help fill the gap created by the lack of studies in geographic areas where Spanish is the official language and it is in contact with English.

Three questions were asked at the onset of this study. The first one was what linguistic factors predict the use of *ser* and *estar + adjective* in the Spanish spoken by Costa Ricans in Limón. The second was what social factors predict the use of *ser* and *estar + adjective* in the Spanish spoken by Costa Ricans in Limón. The last question was whether the pattern of variation in the use of *ser* and *estar* could be considered a change in progress or a stable change. After reviewing previous empirical work done on this topic; four methodological issues were of relevance to the design of the present work. These issues are: (1) the importance of a proper power analysis during the design stages, or the probability of correctly rejecting the null hypothesis when it is false in the population, (2) the analysis of the structure of the data to select the proper statistical analysis to be used, (3) polychotomization, or categorization, of continuous variables and its effect on the power of the study, and (4) the determination of sample size to achieve an adequate level of power. The present work introduces a methodology to determine sample sizes based on power for the study of sociolinguistic data and it provides the field of sociolinguistics with a modification to the variable rule analysis that allows us to see how social predictors help explain the variance, from a statistical point of view. Differences due to social factors are obscured when the data is aggregated; therefore, a
multilevel analysis in the context of logistic regression is suggested as a more powerful method for the prediction of the influence of social predictors on linguistic phenomena.

Five predictors were found to be statistically significant. These predictors were experience with the referent, adverb, subject, resultant state and adjective class. Three of these predictors, experience with the referent, resultant state and adjective class, have been shown to be very strong in the prediction of estar in different social and geographical contexts. However, these predictors are only strong predictors of estar when they are in conjunction with predictors such as predicate reading, susceptibility to change, and gradience. It is determined that variation of copula choice in Limonese Spanish is first and foremost a syntactic phenomenon constraint by discursive and pragmatic features in accord with previous studies of copula choice in Spanish. This study shows that two varieties of Spanish can live in the same geographical area and be constrained by different social and linguistic factors. It also shows how contact with formal education, levels of bilingualism, and gender help explain variation of a syntactic structure. Monolingual Spanish behaves differently than bilingual Spanish because of access to formal education, or lack thereof, and levels of bilingualism. The extension of estar seems to be stable in the monolingual variety of Spanish while it is still ongoing in the bilingual variety of Spanish with younger speakers approximating the use of monolingual speakers.
# TABLE OF CONTENTS

Chapter 1 ........................................................................................................................................ 1  
  Introduction ............................................................................................................................... 1  
Chapter 2 ........................................................................................................................................ 7  
  Theoretical Background ............................................................................................................. 7  
  Introduction ............................................................................................................................... 7  
  Language Change and Language Contact ............................................................................. 8  
  Costa Rica: A Multilingual Society ......................................................................................... 9  
  The Case of the Province of Limón ....................................................................................... 13  
    Socio-historical profile ......................................................................................................... 13  
    Sociolinguistic Profile ........................................................................................................... 15  
  *Ser* and *Estar* ..................................................................................................................... 18  
    Definition of Copular Verbs ................................................................................................. 20  
    Variation and its implication in theoretical accounts on copula choice ......................... 20  
  Early theoretical accounts ....................................................................................................... 20  
  Modern theoretical accounts ................................................................................................. 24  
    Semantics of the verbs and speaker intent .......................................................................... 25  
  The adjective and its importance in the classification of *ser* and *estar* ....................... 29  
  Previous Work on the variation of the structure *ser* and *est* + adjective .................... 44  
  Spanish in Contact with English ............................................................................................. 45  
  Monolingual Spanish ............................................................................................................... 50  
  Spanish in contact with other languages ............................................................................. 57  
  Methodological Issues ............................................................................................................. 61  
  Conclusion ................................................................................................................................. 64  
Chapter 3 ........................................................................................................................................ 67  
  Methodological Issues Addressed ......................................................................................... 67  
  Introduction ............................................................................................................................... 67  
  Applied Linguistic studies and sample size planning ......................................................... 68  
  Importance of a proper power analysis during the design stages .................................... 70
Power ........................................................................................................................................ 71
Importance of power analysis .................................................................................................. 71
Power Analysis ......................................................................................................................... 75
  The statistical test or Statistical Analysis ........................................................................ 76
Significance Criterion (Alpha level) .................................................................................... 76
Effect Size .............................................................................................................................. 79
Timing for power analysis ..................................................................................................... 83
Structure of the data and selection of proper statistical analyses ...................................... 83
  Variable Rule Analysis (VARBRUL) .............................................................................. 84
Statistical Analyses ................................................................................................................ 90
  Logistic Regression ............................................................................................................. 90
    The Fit of the Model ......................................................................................................... 94
  Testing for the significance of the coefficients .................................................................... 97
  Multilevel Logistic Regression ............................................................................................ 100
Polychotomization of continuous variables ........................................................................ 101
Sample Size ............................................................................................................................ 103
  Background ......................................................................................................................... 104
  Importance of proper sample size calculations .............................................................. 105
  Approaches to sample size determination ........................................................................ 107
Determination of Sample size for the study of copula+adjective in Limon, Costa Rica .......... 109
  The statistic test ................................................................................................................ 109
  The parameters .................................................................................................................. 110
  The a priori Monte Carlo Simulations ............................................................................... 111
Chapter 4 .................................................................................................................................. 115
  Methodology ....................................................................................................................... 115
  Introduction .......................................................................................................................... 115
  The Research Questions ..................................................................................................... 115
  The Sample ......................................................................................................................... 116
  Data collection ..................................................................................................................... 118
  Instruments ......................................................................................................................... 118
Informed Consent ................................................................. 118
Background Questionnaire .................................................. 118
Labovian Interviews ............................................................ 121
Procedure ............................................................................. 122
Data coding ........................................................................... 123
Data coding scheme ............................................................. 123
Data Coding Procedure ....................................................... 134
The statistical test and Data analyses ................................. 137
Hypothesis testing ............................................................... 138
Interpretation of the results ................................................. 139
Selecting the variables ......................................................... 144
Chapter 5 ............................................................................ 147
Results ................................................................................ 147
Introduction ......................................................................... 147
Descriptive statistics ........................................................... 148
The sample .......................................................................... 148
General Tendencies .............................................................. 150
Linguistic variables ............................................................. 151
The sentence ...................................................................... 151
The adjective ...................................................................... 157
Social Predictors ................................................................. 162
Selection of Variables ......................................................... 162
Social Variables ................................................................ 163
Linguistic Variables ............................................................ 164
The Models ........................................................................ 166
Formulation of Models ......................................................... 166
Selection of the best fitting model ....................................... 167
Predicate Reading Model 36 .............................................. 170
The multilevel logistic regression ....................................... 170
Linguistic predictors .......................................................... 172
Social Predictors ............................................................... 181
Table 2.1   Number of Speakers per Language Family and Population of Indigenous Peoples of Costa Rica ................................................................. 12
Table 2.2   General Principles on the distinction between *ser* and *estar* ....................... 23
Table 2.3   Copula-adjective combination according to adjective type and reading for both animate and inanimate subjects (adapted from Clements, 2006) ............... 43
Table 2.4   Summary of previous studies on copula choice ................................................. 63
Table 3.1   Effect Size Indices according to Statistical Test ................................................. 82
Table 3.2  Power Calculations for the Omnibus Effect of the Saturated Model ................. 112
Table 3.3    Power Calculations for the Targeted Effect of the Variable Gradiency .......... 113
Table 4.1   Distribution of Speakers by Social Characteristics ........................................ 117
Table 4.2  Coding of Independent Linguistic Variables at the Sentence Level for the Studies of Copula Choice: The Sentence ........................................ 124
Table 4.3  Coding of Independent Linguistic Variables at the Sentence Level for the Study of Copula Choice: Adjectives ......................................... 130
Table 4.4  Coding of Independent Social Variables for the Study of Copula Choice .... 132
Table 4.5  Transcription Code ......................................................................................... 135
Table 4.6  Evidence of Reliability of the Code: 3 Coders .................................................... 137
Table 4.7  Example of "lm()" outcome in R: .................................................................. 140
Table 5.1  Distribution of *ser* and *estar* according to the type of context .................... 149
Table 5.2  Distribution of *ser* and *estar* according to the predictor *predicate reading* .. 151
Table 5.3  Distribution of *ser* and *estar* according to the predictor *susceptibility to change* ......................................................................................... 152
Table 5.4  Distribution of *ser* and *estar* according to the predictor *experience with the referent* ....................................................................................... 153
Table 5.5  Distribution of *ser* and *estar* according to the predictor *adverb* ................ 155
Table 5.6  Distribution of *ser* and *estar* according to the predictor *subject* ................. 156
Table 5.7  Distribution of *ser* and *estar* according to the predictor *resultant state* ...... 157
Table 5.8  Distribution of *ser* and *estar* according to the predictor *adjective class* ...... 158
Table 5.9  Distribution of *ser* and *estar* according to the factor group *underlying structure* ......................................................................................... 160
Table 5.10 Distribution of *ser* and *estar* according to the factor group *gradiency* ...... 161
Table 5.11 Predictive value of individual predictors.................................................. 165
Table 5.12 Number of Possible Models by First Variable Included.......................... 167
Table 5.13 Model with the Lowest Deviance with their respective significant predictors.................................................................................................................. 168
Table 5.14 Multilevel Logistic Regression of predicate reading Model 36 .............. 173
Table 5.15 Interpretation of level 2 parameters of predicate reading Model 36.......... 182
LIST OF FIGURES

Figure 2.1. Linguistic Map of Costa Rica (Gordon, 2005) and area of interest .............. 11
Figure 2.2. Aspectual Marking of ser (Schmitt and Miller, 2007) .................................. 39
Figure 3.1. The Probability of Error in Statistical Significance Testing of Treatment (T) versus Control (C) Group Differences ................................................................. 78
Figure 3.2. Two by two conceptualization of possible scenarios when statistical power is crossed with statistical accuracy ................................................................. 108
Figure 3.3. Sample size calculations for the omnibus effect of the saturated model...... 112
Figure 3.4. Sample Size Calculations for the Targeted Effect "Gradiency" ................. 113
Figure 4.1. Background Questionnaire ........................................................................... 120
Figure 4.2. Limón Corpus Labovian Interview Outline ............................................... 121
Figure 4.3. Visual explanation of the outcome for the predictor underlying structure .. 142
Figure 4.4. Visual explanation of the outcome for the predictor gradiency .............. 142
Figure 5.1. Distribution of ser and estar according to predictor predicate reading ...... 152
Figure 5.2. Distribution of ser and estar according to the predictor susceptibility to change ................................................................. 153
Figure 5.3. Distribution of ser and estar according to the predictor experience with the referent ................................................................. 154
Figure 5.4. Distribution of ser and estar according to the predictor adverb ................. 155
Figure 5.5. Distribution of ser and estar according to the predictor subject .............. 156
Figure 5.6. Distribution of ser and estar according to the predictor resultant state ...... 157
Figure 5.7. Distribution of ser and estar according to the predictor adjective class ..... 159
Figure 5.8. Distribution of ser and estar according to the factor group underlying structure ................................................................. 160
Figure 5.9. Distribution of ser and estar according to the factor group gradiency .... 161
Figure 5.10. Multilevel logistic regression of the predictor predicate reading .......... 175
Figure 5.11. Multilevel logistic regression of the predictor susceptibility to change .... 176
Figure 5.12. Multilevel logistic regression of the predictor experience with the referent ................................................................. 177
Figure 5.13. Multilevel logistic regression of the predictor adverb. ............................ 178
Figure 5.14. Multilevel logistic regression of the predictor subject ........................................ 178
Figure 5.15. Multilevel logistic regression of the predictor resultant state ............................... 179
Figure 5.16. Multilevel logistic regression of the predictor adjective class .............................. 180
Figure 5.17. Multilevel logistic regression of the predictor gradiency ......................................... 181
Figure 5.18. Production of estar by age .................................................................................... 183
Figure 5.19. Confidence intervals of the intercept by age .............................................................. 184
Figure 5.20. Production of ser and estar by gender ................................................................. 185
Figure 5.21. Production of ser and estar by level of education ...................................................... 186
Figure 5.22. Production of ser and estar by bilingualism ............................................................. 187
Figure 5.23. Production of ser and estar by bilingualism ............................................................. 187
Figure 6.1. Distribution of predictor resultant state by gender and bilingualism ...................... 196
Figure 6.2. Distribution of gender by level of education .............................................................. 200
Figure 6.3. Distribution of speakers by bilingualism and level of education ......................... 201
Figure 6.4. Distribution of uses of estar by gender and level of education .............................. 202
Figure 6.5. Distribution of gender by bilingualism ................................................................. 203
Figure 6.6. Distribution of uses of estar by bilingualism and level of education ..................... 204
Figure 6.7. Distribution of production of estar by gender and level of education among monolinguals ................................................................................................................. 205
Figure 6.8. Distribution of production of estar by gender and level of education among bilinguals ............................................................................................................................................ 205
Figure 6.9. Production of estar by gender, bilingualism and level of education ....................... 206
Figure 6.10. Distribution of bilinguals’ use of estar by age ....................................................... 209
Figure 6.11. Distribution of monolinguals’ use of estar by age .................................................... 209
Chapter 1

Introduction

Various studies of copula (i.e., *ser* and *estar*) choice in Spanish in contact situations, as in among Spanish-English bilinguals, in contexts where Spanish is not an official language (i.e., Southwestern United States) have shown that copula selection is undergoing change. The change from traditional usage is that the copula *estar* is becoming more accepted in an extended pre-adjectival context, especially with adjectives of size, physical appearance, age, and description and evaluation: this change was initially attributed to contact with English (Silva-Corvalán, 1986). This claim has been extended to monolingual Spanish by Gutiérrez (1992, 1994) because he found that monolingual Mexicans from the same social class as those studied by Silva-Corvalán (1986) showed the same behavior with regard to copula choice. Furthermore, Diaz-Campos & Geeslin (in press) found evidence of this phenomenon in an analysis of the spoken Spanish of Caracas, a context where Spanish is the official language and has no contact with English.

Due to this debate, copula selection has become a prolific field in both first language (Diaz-Campos & Geeslin, in press; Guijarro-Fuentes & Geeslin, 2006; Silva-Corvalán, 1986) and second language studies (Geeslin, 1999, 2000a, 2000b, 2001, 2002, 2003). However, previous studies on the Spanish copula have mostly investigated settings where either Spanish is the only language spoken and have been categorized as monolingual (e.g., Caracas, Venezuela), Spanish is not the official language (e.g., Los Angeles), and contexts where Spanish is spoken widely, but English has an important
role in formal education (e.g., Puerto Rico). Another context in which copula choice in Spanish has been investigated is the acquisition of Spanish as a foreign or second language. These studies have been conducted in settings where Spanish is not an official language (e.g., United States). Therefore, the premise of this work is to help fill the gap created by the lack of studies in geographic areas where Spanish is the official language and it is in contact with English.

Variation, as a sociolinguistic term, has been defined as two or more realizations of a common element (Labov, 1974). It is commonly referred to as two or more ways of saying the same thing [my translation] (Silva-Corvalán, 2001). The term was coined to explain phonological phenomena, but scholars have attempted to apply it to the study of syntactic phenomena. Researchers have encountered empirical and analytical problems for the adoption of such a definition to explain this type of phenomena. Lavandera (1978) argues against extending this definition because of the lack of a theoretical framework related to meaning at the syntactic level that could aid quantitative studies of this sort. In other words, she argues that there is no theoretical framework that can explain whether two or more syntactic structures really are two or more ways of conveying the same meaning. Lavandera (1978) proved that the correlation between social and phonological variables is statistically significant, but the correlation between social and syntactic variables is not. This lack of correlation between social and syntactic variables is due to the fact that the differences in frequencies found with syntactic variables may be related to the differences in meaning they carry. She proposes that these variables need to be studied in a special way, a stylistic one. In other words, Lavandera (1978) proposes that the syntactic variable can be analyzed according to the meaning of each variant. Their
meaning could be described as different communication style markers such as more or less assertive, more or less polite, among others. For this type of variable to be a sociolinguistic one, it must comply with the condition that the variants must have social, stylistic or other meaning beyond their referential one. Furthermore, quantification of the variants must allow for relative frequencies of occurrence of certain factors to be the direct indicators of their non-referential meaning (Lavandera, 1978). However, at the time that Lavandera (1978) made this claim, statistical analyses were not powerful enough to test for the influence of social variables with respect to syntactic structures believed to be found in variation.

Silva-Corvalán (2001) expands Lavandera’s (1978) requirements by stating that the nature of the syntactic variation is not analogous to the nature of the phonological variation as we know it. She points out the following differences:

There is less syntactic variation than phonological variation in a single variety of a language. A syntactic variable usually has two variants while the phonological variable has at least three or more.

Syntactic variation is more difficult to study, especially to quantify. This difficulty arises due to the low frequency of contexts where syntactic variation is found and to the difficulty of finding them without eliciting them directly.

Identification and definition of contexts for syntactic variation is more difficult than those for phonological variation.

Syntactic variation has the problem of potential differences in meaning associated with each variant. In other words, there may not be true synonymy.

Silva-Corvalán (2001) states that due to these reasons, it has not been easy to assign social and/or stylistic value to cases of syntactic variation. She states that in most of the cases, syntactic variation seems to be conditioned by syntactic, semantic, pragmatic, and sometimes phonological factors but not social ones. Due to the debate on whether to subscribe strictly to the definition of pure synonymy or not, Sankoff (1988a) proposes the notion of weak complementary distribution. To exemplify this notion, Silva-
Corvalán (2001) uses the case of *ser* and *estar* with predicate adjectives in Morelia, México studied by Gutierrez (1994a). In this study, Gutierreiz (1994a) found the following examples:

1. Las casas están chiquitas, pero bonitas.
2. Las casas son chiquitas, pero bonitas.
   ‘These houses are small, but pretty’

Silva-Corvalán (2001) explains that the analysis of these cases poses two interpretations that support Sankoff’s (1988a) definition of weak complementary distribution. Both forms are either in true synonymy or they have two different functions (Silva-Corvalán, 2001). If the latter case is true, Silva-Corvalán (2001) states, some speakers use number one in one context with one meaning and other speakers use number two in the same context with a different meaning. If that assertion is true, they are not impure complementary distribution. However, if the former case is true, then these two examples are cases of true synonymy and they are found to be true syntactic variation.

The difficulty of determining which analysis best fits the case for *ser* and *estar* makes this phenomenon a good case for Sankoff’s (1988a) notion of weak complementary distribution. Silva-Corvalán (2001) proposes that in such cases the distinction among the linguistic variants is not relevant for the speaker, which puts them in variation. She also states that cases of weak complementary distribution are, often times, signs of a change in progress.

*Ser* and *estar* represent good cases for testing the existence of syntactic variation as we know the term “variation”, i.e., two or more ways of saying the same thing; and it also serves to test whether or not there is a change in progress. Sankoff’s (1988a) definition of weak complementary distribution helps us to explain a syntactic system in which not all its parts are in variation and allows us to expand our notion of syntactic
variation. Silva-Corvalán (1986) states that “the extension of estar in progressives and its frequent association with ‘be’ in these constructions may favor the rapid diffusion of estar in the context of predicate adjectives, where Spanish evidences a slow progress of change independent of any language-specific influence” (p. 604). But she clarifies that there is no evidence that the dual system of copulas is going to collapse regardless of contact with systems with only one copula or the rapid expansion of estar. She states that language contact situations accelerate change, but do not cause it.

I believe that the language situation found in Limón, Costa Rica and the case of copula + adjective serves as a good test case for finding an approach to the definition and study of syntactic variation. By proposing a statistical model for the prediction of the usage of estar to compare modern uses with traditional uses of it, I will be able to determine whether both Spanish copulas convey the same meaning. The language situation found in Limón shows features that allowed me to test the different hypotheses regarding copula + adjective constructions and the extension of estar. First, it represents a geographic area where Spanish is learned and spoken alongside English. English has a status of a minority language or language of a particular social group (i.e., Afro-Costa Ricans). Second, Spanish is the medium of instruction at all levels of education, whereas English is taught as a foreign language. Third, it represents a setting that has not been studied before where a monolingual variety of Spanish abides with a bilingual variety of Spanish. By being able to statistically predict the modern usage of estar, another hypothesis can be tested: the contact hypothesis related to copula choice. So, if contact accelerates change and change is slow in a monolingual variety; a place where these two varieties reside together would make a good test case for these two hypotheses. Although
there is enough evidence against the former (Geeslin & Guijarro-Fuentes, 2008; Ortíz-
López, 2000), more evidence is needed to reject or accept the hypothesis that contact
accelerates change.

The chapters of this dissertation contain a theoretical background that includes a
contextualization of the contact situation in Limón, a socio-historical account of the
linguistic situation, and a thorough discussion of copula choice in Spanish. Previous
studies of copula choice are presented with two objectives in mind. The first objective is
to select factors to be used in the analyses. The second is to look at the procedure for
sample size determination that was followed in order to help with the determination of
the sample size for this project. Chapter 3 contains methodological issues that were
addressed during the design stages of this study. Chapter 4 outlines the methodology
followed for the present study. Chapter 5 presents the results of the analyses. And,
Chapter 6 outlines the discussion and respective conclusions of this study.
Introduction

The present chapter contains the theoretical foundations for the study of copula choice in the Spanish spoken in Limón, Costa Rica. The main objective for the presentation of these foundations is twofold. First, it provided me with the general theoretical and empirical approaches to the study of copula choice. Second, it supplied me with a set of factors for the construction of a statistical model for the prediction of \textit{estar}. Therefore, studies were selected because of their contributions to the general knowledge regarding copula choice in Spanish and to search for possible predictors for the creation of a model for the prediction of the use of \textit{estar} in the Spanish of Limón. The chapter is divided into six sections not including this introduction. The first section contains a definition of language contact and language change. The second contains a socio-historic and sociolinguistic background of Costa Rica and Limón. The third contains theoretical accounts related to \textit{ser} and \textit{estar}. The fourth contains previous work on the variation of the structure \textit{ser} and \textit{estar + adjective} in Spanish. The fifth contains a list of methodological issues that arose from the previous research conducted on copula choice in Spanish. These issues were addressed in the design stages of this study and are also discussed in Chapter 3. Finally, the last section is the concluding remarks for this chapter.
Language Change and Language Contact

Thomason (2001) defines language contact as ‘the use of more than one language in the same place at the same time’ (p. 1). In Limón, we find asymmetrical bilingualism in which, according to Thomason (2001), a subordinate bilingual group (i.e., Limonese English speakers) is believed to be shifting to the language of a monolinguistic dominant group (i.e., Costa Rican Spanish). Due to this situation in Limón, language changes are present in two directions. The first is the influence of Limonese Spanish on Limonese English, and the second the influence of Limonese English on Limonese Spanish. The latter represents a context in which copula choice has not been investigated and is the main aim of this study.

In order to study copula choice in Limón, this dissertation is set within the variationist framework. Two approaches are taken into account. The first is Labov’s (1974) mechanisms to account for language change and the second is Díaz-Campos and Geeslin’s (in press) approach to the analysis of copula choice.

Labov (1974) puts forward three steps for the study of linguistic change in progress. These three steps are the following:

1. The transition problem is to find the route by which one stage of a linguistic change evolved from an earlier stage.
2. The embedding problem is to find the continuous matrix of social and linguistic behavior in which the linguistic change is carried.
3. The evaluation problem is to find the subjective (or latent) correlates of the objective (or manifest) changes which have been observed. (p. 161-162)

Labov (1974) also proposes that the value of an explanation increases in relation to its generality, but only to the extent that it rests upon a foundation of reliable and reproducible evidence. To search for this type of evidence, he states that the simplest data
that can establish the existence of a linguistic change is a set of observations of two successive generations of speakers. These generations should be of comparable social characteristics that represent stages in the evolution of the same speech community.

Finally, Labov (1974) observes that solutions to transition problems (due to the observation of two generations) depend upon close analysis of the distribution of linguistic forms along the dimensions formed by the age groups of the population (i.e., apparent time). Apparent time is a construct that states that when social and stylistic factors are held constant, linguistic differences among different generations of a population would mirror actual diachronic changes in language (Bailey, 2002). Following this assumption, the present dissertation is based on the concept of apparent time and was conducted by taking into account the different generations of bilingual speakers of Limonese English and the varieties of Spanish found in Costa Rica; however, age will be regarded as a continuous rather than a grouping variable. An explanation of this decision is given in Chapter 3 in the subsection called ‘Polychotomization of continuous variables’.

Because my interest lies in the syntax of oral interviews with regard to the study of a particular structure (i.e., copula + adjective), I present the theoretical accounts of Spanish copula choice use and variation in several varieties of Spanish and on copula use in Limonese English. The reader is directed to the works of Herzfeld (2004) and Winkler (1998) for a full account of the description of Limonese English and to Arroyo (1971) for a perspective of the Spanish as spoken by Costa Ricans.

Costa Rica: A Multilingual Society

Even though Costa Rica has been traditionally categorized as a monolingual society, the linguistic situation in this country is not different from most multilingual
societies in the world (Aguilar-Sánchez, 2005a, 2005b). Due to the fact that Costa Rica’s official language is, according to the present Costa Rican constitution, Spanish, other languages spoken in the country are neglected for the most part. Linguistic phenomena within Costa Rica have not yet been investigated from a multilingual perspective. This being the case, it is important to highlight that the language situation in Costa Rica is indeed more complex than what is normally believed, because more than three different languages are spoken in the country. As a result, Costa Rica’s linguistic situation cannot be categorized as that of a monolingual society. Like most of the countries in Central America, Costa Rica is a country with native English speakers in a small percentage (about 2 percent) of its population (Aguilar-Sánchez, 2005a, 2005b).

Due to the popular belief that the linguistic situation in Limón is an isolated phenomenon particular to this small region of the country, the majority of studies on this variety of English have focused on the description of linguistic and/or cultural phenomena within the field of Creoles (Herzfeld, 1980, 1994, 2003). However, the impact of this English on the Spanish of the country and specifically on the Spanish of the province of Limón has not yet been investigated. This lack of scholarly interest in the impact of multilingualism in the area of Limón and/or of Costa Rica and its linguistic diversity, in general, constitutes a primary reason for undertaking this study. The study of such phenomena at a micro level will be applicable to the study of these phenomena at a macro level.

Many of the languages spoken in Costa Rica have been neglected and many of them can be considered to be near extinction. Within its small territory, we find seven
Amerindian languages (6 alive and 1 extinct), six transplanted languages, a sign language, and the official language (see Figure 2.1).

According to Gordon (2005), the Amerindian languages include languages from the Chibchan/Talamanca, the Chibchan/Rama, and the Oto-Manguean/Chiapanec-Mangue families (See Table 2.1 for speaker statistics).

Similar to the situation of the Amerindian languages, Costa Rica’s linguistic profile is enhanced by the presence of six transplanted non-Amerindian languages. Of these, English is found in Costa Rica as a result of work-related diaspora, more specifically as a result of the construction of the railroad to the Atlantic by the United Fruit Company in the late 19th and early 20th centuries (Aguilar-Sánchez, 2005a, 2005b; Herzfeld, 2004; Meléndez Chaverri & Duncan, 1974; Winkler, 1998; Winkler & Obeng, 2000).
Table 2.1

<table>
<thead>
<tr>
<th>Language Family</th>
<th>People</th>
<th>Language</th>
<th>Speakers</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chibchan/Talamanca</td>
<td>Borucas</td>
<td>Boruca</td>
<td>35</td>
<td>1000</td>
</tr>
<tr>
<td>Chibchan/Talamanca</td>
<td>Bribris</td>
<td>Bribri</td>
<td>11000</td>
<td>12172</td>
</tr>
<tr>
<td>Chibchan/Talamanca</td>
<td>Cabécares</td>
<td>Cabécar</td>
<td>8840</td>
<td>9308</td>
</tr>
<tr>
<td>Chibchan/Rama</td>
<td>Malékus</td>
<td>Maléku Jaíka</td>
<td>750</td>
<td>1074</td>
</tr>
<tr>
<td>Chibchan/Talamanca</td>
<td>Teribes</td>
<td>Teribe</td>
<td>5</td>
<td>300</td>
</tr>
<tr>
<td>Extinct</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oto-Manguean/Chipanec-Mangue</td>
<td>Chorotega</td>
<td>Chorotega</td>
<td>0</td>
<td>795</td>
</tr>
</tbody>
</table>

This phenomenon is also strengthened by the contemporary immigration of retirees from English-speaking countries (Aguilar-Sánchez, 2005b). Puga (2001) states that it is very difficult to determine the exact number of foreign retirees in Costa Rica due to the different immigration categories of residency types that can be found throughout the immigration and legal system. However, she presents two numbers for 1998. The first one is the number provided by the Costa Rican Institute of Tourism (ICT Spanish abbreviation) of 9,370 and the second is the number provided by the Association of Residents of Costa Rica (ARCR) of 20,000. These numbers represent a per capita number higher than that of Mexico given that the population in Costa Rica at the time was approximately three million eight hundred thousand (Puga, 2001).

Work-related diaspora brought English to Limón. About 55,100 Afro-Costa Ricans, mainly in this province, speak Limonese English at the present time. This diaspora also introduced two Chinese dialects: Ngäbere and Yue. These varieties are spoken by approximately 5,092 and 4,500 people respectively. In addition, with
contemporary immigration came Plautdietsch, Basque, and Eastern Yiddish. Plautdietsch, a Germanic language, is spoken by about 100 people in the northern area of the country. The numbers of the latter two are not documented.

In this work, I deal with the phenomenon of copula choice in Limón, Costa Rica. A brief history of the linguistic situation of this region of Costa Rica is presented here to help contextualize the current study of copula choice in Spanish in a context where English is not an official language.

The Case of the Province of Limón
Socio-historical profile

Situated on the Atlantic coast of Costa Rica, the province of Limón represents one of the most complex language settings in Costa Rica. English was introduced to the territory by the construction of the railroad to the Atlantic. It was built by the United Fruit Company, an enterprise attracted by the blossoming of banana plantations in the region in the late 19th and early 20th centuries. Workers were brought from Jamaica and other islands of the Caribbean not only for the construction of the railroad but also because of the opportunity to work on the plantations (Aguilar-Sánchez, 2005a, 2005b; Herzfeld, Carrington, Craig, & Dandare, 1983; Meléndez Chaverri & Duncan, 1974; Palmer, 1986; Purcell, 1993; Viales Hurtado, 1998; Winkler, 1998). English has been the main language of intra-racial communication and a representative icon of the Afro-Caribbean ancestry for the population of the region (Aguilar-Sánchez, 2005a, 2005b; Viales Hurtado, 1998).

Maintenance of English was predominant among inhabitants of Limón in the early years of the 20th century. By 1927 the majority of the Jamaican workers in Limón had achieved at least a primary education level. The literacy rate for English of people
nine years and older was about 77.4 percent (Viales Hurtado, 1998). This high rate of literacy was attributed to the great number of private English schools found in the region. According to the Dirección General de Estadísticas y Censos [Census Bureau] (1927 in Viales Hurtado, 1998) there were thirty-three private schools, most of them of religious origin and their medium of instruction was mainly English. Most schools disappeared after the region became an important economic area for the country. The disappearance was caused, partly, by the implementation of national education policies that established Spanish as the language of education in the region. Today, the literacy rate for Spanish in the country is about ninety-three percent. Limón’s literacy rate for Spanish is comparable due to efforts from the government to provide formal education to every Costa Rican. As mentioned previously, this literacy rate is based on the knowledge of Spanish, which is Costa Rica’s medium of instruction by law. Contemporary English literacy rate in Limón cannot be presented here because of the lack of statistical information. Several private schools have English as the medium of instruction and most public schools teach English as a foreign language from first grade (Aguilar-Sánchez, 2005b); therefore, the literacy rate in English varies according to the type of education (i.e., private or public) that the residents of Limón receive.

Another type of immigration was triggered by the railroad construction and the growing number of plantations in the area. This immigration wave came from within Costa Rica. Several factors prompted the migration to the region among which we find the strong desire by the government to decrease its economic dependency on one crop (i.e., coffee). First, the government implemented policies that served as the beginning of the banana plantation era. Second, the government gave land and privileges to migrants
as an incentive to move to the area. Third, the railroad construction created new economic regions where people eventually settled. Finally, the sustained population growth due to the coffee plantations in other parts of Costa Rica forced people to migrate to this new land of opportunities (i.e., Limón) (Viales Hurtado, 1998). Thus, by the mid 20th century the population of Limón had increased greatly due to this internal migration. As a result, Spanish took over most of the domains that English once held.

Spanish and English are not the only languages that are present in the province of Limón. In this area, we also find Chinese and Bribri languages that were mentioned in the previous section. I will abstain from detailing their development because the main purpose of the current study is related to the contact situation between Spanish and English.

Sociolinguistic Profile

In the main city of Limón, Puerto Limón, as well as in the other cities along the railroad in the Atlantic region of Costa Rica, about seventy-six percent of Afro-Costa Ricans speak English as a first language and Spanish as a second language (Winkler, 1998). In this region, speakers of English were believed to be moving towards adopting Spanish, the dominant language, in all social domains, a phenomenon that Thomason (2001) calls asymmetrical bilingualism. This situation, however, is changing due to the recent growth of the importance of English as a means of social mobility that has surfaced in Costa Rica in the last two decades (Aguilar-Sánchez, 2005a). Aguilar-Sánchez (2005a) found that Limonese English speakers have become aware of the importance of formal education in English resulting in a movement towards a total bilingualism with the aid of English and Spanish education. Bilingual (English/Spanish) schools have sprouted as a result of the recent development of the tourism industry and
the implementation of English as the first foreign language at all levels of formal schooling.

Due to the constant change of importance between these two languages throughout the twentieth century, it is necessary to describe the predominant language situation and language use by residents of Limón in the last one hundred years. In the past, Limonese English was reserved for intra-racial relationships, religion, use at home, and for working with tourists. On the other hand, for speakers of Limonese English, Spanish had been used in almost all domains of everyday interaction with other Costa Ricans, for the government, and more recently for religious purposes.

Primary and Secondary education have been predominantly taught in Spanish. However, schooling has undergone changes since the establishment of the province of Limón as an economic region in the early 20th century. Initially, all instruction was done in English; teachers were brought from Jamaica to teach in the region. Then, with the intra-Costa Rican migration to Limón in search of land and opportunity, the government expanded its educational policies to include the entire province. This expansion triggered a shift from English medium schools to Spanish medium schools for everyone. Finally, more recently, the need for bilingual education for social mobility has prompted the implementation of bilingual schools. In formal education, English is being taught as a foreign language instead of English for native speakers.

Winkler (1998) asserts that in the last thirty years, the great majority of Limonese English speakers have become bilingual in Spanish, but that Spanish has not altogether replaced Limonese English in the community. English is still used for intra-racial and family communications. It is also used for religious services at multiple churches. More
recently, English is being used to communicate with tourists that visit the region. In addition, Winkler (1998) states that there has also been an increase in the influence of Standard English in this community. More recently, Aguilar-Sánchez (2005a) found in his survey of attitudes and demographics that Limonese English is going through different processes that induce change towards standardization. These processes include a higher rate of Limonese speakers searching for formal education in English, which is caused both by positive attitudes and educational policies towards standardized varieties of English such as American and British. Additionally, blossoming of the tourism industry brings Limonese people in contact with English speakers from other countries (e.g., USA, Britain, and Canada, among others).

These changes in language policies and language attitudes have created a complex linguistic context in which we can find people who grew up in bilingual environments from birth, bilinguals that learned their second language at school, and descendants of speakers of English that are monolingual in Spanish (i.e., they have shifted to the dominant language). This phenomenon is of special interest for a language contact researcher because observing changes in the social status of a language over short periods of time might help us to understand patterns of language use and language change. Therefore, it makes it possible to document phenomena that otherwise will take several hundreds of years to document (Thomason, 2001).

In her work, Winkler (1998) suggests that Limonese English makes a good test case for Thomason’s and Kaufman’s (1988) model. This model states that intense structural borrowing will only be observed after perhaps a couple hundred years of intense cultural contact which Winkler (1998) found to be true. Limonese English is a
case in which changes have been documented since the very beginning of contact with Spanish which, in turn, permits language change, maintenance, and shift to be studied and evaluated as they occur. In Limonese English for predicative constructions has different realizations (Winkler, in progress). These realizations include “a”, “is”, “wuz”, “were”, and zero copula. Winkler (in progress) states that these are just stylistic variants for many speakers and do not consistently reflect tense or person differences. In fact, there is a continuum of overlapping speech varieties without a break, stretching from the most basilectal forms of Limonese Creole to Standard English (Herzfeld, 1978; Purcell, 1993; Winkler, 1998, in progress). The fact that Limonese English presents a complex copular verb system and that their use may be traceable to sociolinguistic factors such as level of education makes a researcher of Spanish ask what seems to be the most obvious question: is there any influence of Limonese English on Limonese Spanish with regard to copula choice and copula use?

**Ser and Estar**

This dissertation particularly focuses on a specific linguistic phenomenon: copula use in Limonese Spanish. It contributes to the general literature in language variation by providing an account of *ser* and *estar* by using Diaz-Campos and Geeslin’s (in press) variables and coding techniques of *copula choice* in the context of *copula + adjective* in a language contact environment where Spanish is the predominant language. It also helps to understand if there is influence from the English of this region, in the copula system of this particular Spanish.

Over the years three different areas of linguistic inquiry have attempted to explain the differences and variation found between Spanish copular verbs *ser* and *estar*. All three areas of knowledge have added key points on the undergoing discussion and
understanding of these phenomena. These areas are prescriptive/descriptive grammar as early as the late 19th century, theoretical linguistics and sociolinguistics. Furthermore, at least four different approaches have been taken. The first one is related to the semantics of the two verbs, the second to the predicate type, the third related to discourse-pragmatics, and the fourth related to usage.

In the following subsections, I discuss relevant early and modern approaches to the differentiation between both verbs, challenges to these definitions derived from theoretical accounts of copula choice, and sociolinguistic studies on copula use by native speakers of Spanish. The studies discussed below informed the design of the present work. Most of the discussion is centered on the context of copula + adjective; however, when deemed important other contexts are also discussed. The search for an explanation of the use of ser and estar is not new in the field of Hispanic linguistics. This discussion has moved from definitions pertaining to the semantics of each verb, in an attempt to narrow it to a formal rule, to the analysis, with regard to speaker’s intention, of each verb in the linguistic context in which it appears. However, variation and usage have posed a challenge for grammarians in achieving the goal of formulating a single rule to explain copula choice and this has led us to the study of the use of these verbs as a sociolinguistic phenomenon.

At this point it is important to clarify that some grammarians use the predicate to study this type of phenomena. For some, it includes the components following the verb phrase (i.e., adjectives, nouns, phrases) (Fogsgaard, 2000; Porroche Ballesteros, 1988). For others it includes the whole verb phrase (i.e., verb + complement) (Smith, 1997; Vendler, 1967). Others use the sentence to study such phenomena. In such an approach,
the analysis includes the verb and its complements (i.e., subject, verb, and object, and adjuncts) (Verkuyl, 1993). In this dissertation, I will use this last approach (Verkuyl, 1993, 2001).

Definition of Copular Verbs

Porroche Ballesteros (1988) defines the copular structures as those in which predicates are attributes (i.e., a nominal element: noun or adjective). She states that a copular verb is nothing but a carrier for the morphemes of number, person, tense, mode, and aspect. Fogsgaard (2000) defines the function of a copular verb as that of providing support for predicates of time, manner, aspect, person and number. Winkler (in progress) proposes three possible realizations of the copula: equative (or identificational), predicative, and locative. She explains that the equative is used to link the subject and an object (She is the teacher), the predicative before predicative adjectives (I am cold), and the locative to denote the location of an object (It is here). In Limonese English, zero copula constructions are also possible, especially in locative constructions.

Variation and its implication in theoretical accounts on copula choice

In this subsection, I present some of the most relevant discussions on the semantics of ser and estar and on their attempt to deal with variation and use of these two verbs. The works presented here have helped the development of linguistic analysis and understanding of copula variation and use in Spanish and are part of the foundation for the present work.

Early theoretical accounts

Early theoretical accounts on copula distinction that deal with variation date back to the late 19th and early 20th centuries. At the time, the discussion hovered around the distinction between permanency and temporariness. As a result of that discussion, a clear-
cut distinction was set forth by Bello and Cuervo (1911, 1916). They defined *ser* as expressing permanent and essential conditions and *estar* non-permanent or transitory ones and they regarded any variations as idiosyncrasies or idiomatic usages of the language.

One of the earliest works that argues against this clear-cut distinction in search of an explanation to the many uses that could not be categorized by this dichotomy is that of Andrade (1919). He proposes that there are two ways in which one can search for an explanation to the phenomenon found in *ser* and *estar*. The first one is to search for historical developments from Latin or Old Spanish and the other is by looking at modern uses of these two verbs. In his work, Andrade (1919) dismisses any relationship between the modern uses of *ser* and *estar* and Latin *sum*, because *sum* performed as both modern Spanish copulas and *sto* implied relatively permanent elements rather than transitory ones. He also discards any relationship with old Spanish because *ser* was used more like Latin *sum* (e.g., Toda mi facienda et la mi vida es a grand peligro de se perder ‘My whole land and life is in great danger of being lost’ [my translation]; (Juan Manuel: Conde Lucanor in Andrade, 1919). He also proposes that in order to understand the differences between *ser* and *estar*, we must look for the modern usages of these verbs. Andrade (1919) recommends using the extension of *estar* with respect to the restriction of the original meaning of *ser* as a point of departure to the analysis and explanation of any difference between these two verbs. In his analysis, Andrade (1919) proposes that the expression of immediate perceptions may be one of the latest developments of *estar*. He suggests that with a remark at the table of ‘¡Qué buena está esta sopa! ‘This soup is really good!’ [my translation] one does not mean that the soup is temporarily good or that
the soup taste is an accidental quality of soups; on the contrary, it simply means that the soup tastes good. He concludes that \textit{estar} is associated with the characteristic feelings which attend immediate perceptions and their representations, while \textit{ser} is likewise related to concepts and judgments. He supports his distinction by stating that in judging affective elements our point of view must always be the purpose of the speaker or writer and that in sentences of the type \textit{El libro es rojo} [the book is red] the purpose is to express a judgment even though the predicate is the name of a sensation (i.e., the perception of a color). Andrade’s (1919) work resembles contemporary discourse-based work on the study of variation of the Spanish copulas. Because of his approach of taking into account the purpose of the speaker or writer, his work represents one of the earliest essays that tries to explain the variation found among Spanish copulas from a non-traditional perspective. Although his intention was merely pedagogical and unrelated to modern discourse analysis or the study of variation, as we know it, his approach serves as a point of departure to help us understand how the presence of variation has made scholars revise the earliest approaches to this phenomenon.

Morley (1925), in his quest to find a formal rule or set of rules to explain the usages of \textit{ser} and \textit{estar} and in a response to Andrade’s (1919) approach of feelings/judgment, proposes that ‘[t]he distinction in use between \textit{ser} and \textit{estar} constitutes one of the thorniest points in Spanish grammar’ (p. 541). He believed that it was difficult to reduce this system to one formal rule and suggests that the use of \textit{ser} and \textit{estar}, instead, involves four antitheses that are sometimes mutually exclusive (See Table 2.2).
Table 2.2

General Principles on the distinction between *ser* and *estar*

<table>
<thead>
<tr>
<th>Principle</th>
<th><em>Ser</em></th>
<th><em>Estar</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Duration, inherence</td>
<td>Transiency, accidentally</td>
<td></td>
</tr>
<tr>
<td>II. Mere existence</td>
<td>Definite location in space</td>
<td></td>
</tr>
<tr>
<td>III. With past participles, act.</td>
<td>State (with past participles or adjectives)</td>
<td></td>
</tr>
<tr>
<td>IV. Mental concept</td>
<td>Sensory perception</td>
<td></td>
</tr>
</tbody>
</table>

In his proposal, Morley (1925) identifies only one context in which there are no exceptions to his principles. This particular context is when *ser* is used to express mere existence. For all other contexts, he found many exceptions to his principles. Furthermore, for all of them *ser* and *estar* were found sometimes with no difference in meaning (i.e., in true synonymy). Morley (1925) found that the context most difficult to explain was when both copulas are found with adjectives, adjectival phrases, or verbal derivates that function as adjectives, each of which is a focus of this dissertation.

Regarding variation present in his data, Morley (1925) found that in some uses of *ser* and *estar* two ideas coalesce especially with adjectives. He also stated that some examples are quite idiomatic and without any rational explanation. To explain the impossibility of looking for systematicity in his data, Morley (1925) uses the metaphor that language is a creature of feeling, not of reason, which resembles the concept that languages are in constant change and alive, one of the underpinnings of this dissertation.

Morley (1925) concludes that no one underlying principle suffices to explain all of the uses of *ser* and *estar* and that, when analyzing each of his examples, usually the contrast is between durative and transitory, or between act and state. In addition, *estar* keeps its proper and original sense of location. Morley (1925) also concludes that the conceptual and perceptual principles enter in as secondary connotations rather than
primary ones. Morley (1925) agrees with Andrade (1919) that emphasis is laid now on the point of view of the speaker, at the moment of speech or at another time. He extends his conclusion stating that since at least the seventeenth century, *estar* has invaded the field once held by *ser*, and there is reason to believe that the invasion has not entirely stopped.

Because of the difficulty in the late XIX and early XX centuries, in finding a formal and objective definition underlying the nature and usage of these verbs, linguists moved from prescriptive/descriptive accounts to approaches where the intent of the speaker is taken into account (i.e., what is known as modern discourse/pragmatic analysis). New approaches to the analyses of these two verbs and their uses developed following this line of thought; however, the search for a formal rule to explain these two verbs did not end and is present in some of these new approaches.

**Modern theoretical accounts**

Modern theoretical accounts for the classification of *ser* and *estar* can be divided into four major approaches relevant to the present work from three different theoretical frameworks. The first approach is the one dealing with the semantics of the verbs and speaker’s intent, the second with the type of predicate they are found with, the third with the subject and its role, and fourth with the role of adverbs as temporal/aspectual anchors. The three theoretical frameworks used are functional approaches to syntax, generative syntax, and variationist sociolinguistics.

Some more recent authors work in one, two or all of these areas. Each of the following sections is designed to present accounts that try to explain the distinction between *ser* and *estar* using one of these approaches. If the authors explain the phenomenon using more than one of these approaches, their work is divided and
presented in the appropriate sections. The studies were chosen for their relevance to the present study. Their theoretical framework did not pose major influence in their selection. I searched for approaches that would enable me to construct a statistical model for the prediction of *estar* in Spanish; therefore, if a study provides a possible predictor of *ser* and *estar*, it is presented here. Because this area of research is a very prolific one, not all studies are presented here and not all variables used in those studies were selected for the present work.

**Semantics of the verbs and speaker intent**

Bull (1942) adopts a new approach of norm-versus-change instead of the traditional inherent-accidental/permanent-temporary one. He proposes that besides the semantics of the adjective, a point discussed in the following section, the fundamental differentiation between these two verbs with predicate adjectives is based on change. He points out that it is important to focus on the subjective rather than on the objective reality. He proposes that state, condition, or quality, as objective entities, be replaced by the subjective concept of these entities. He defends the fact that the possibility of a subjective change of concept triggered by a change in attitude unrelated to objective reality explains innumerable uses of the two verbs.

Bolinger (1944) argues that Bull’s (1942) idea of change is more than a terminological shift because even though Moellering (1944), is right in pointing out the philosophical background of essence and accident, and in combating the idea that accident here signifies ‘mishap’, essence-inherence and accident are objective concepts and they have to do with the being of the thing. Bolinger (1944) extends his support for Bull’s (1942) approach stating that when one may choose between *Juan es calvo* and *Juan está calvo* ‘Juan is bald’, both in reference to the same immutable fact; one is either
contradicting oneself or there must be some factors present which are neither essence nor accident. Bolinger (1944) assures the reader that this factor is the speaker’s subjective concept of his/her reality. He emphasizes that objectivity or the search for a formal rule to explain this phenomenon must be used as a final test for *ser* and *estar*, not the first test, because objective permanence influences the average human being to form a subjective concept that will call for *ser*. He continues by saying that if Juan is repeatedly observed to be bald, the speaker can hardly avoid the conclusion that *Juan es calvo*. The author proposes that what immediately underlies the choice of verb is the objective fact that always works through the concept which in turn makes the statement fallible. Because it is a fallible statement, he argues, the most one can say of the sentence *Juan es calvo* is that baldness constitutes the norm about Juan for the speaker which becomes an objective evaluation of Juan that needs to be verbalized; therefore, the use of *ser*.

De Mello (1979), on the other hand, proposes that a clear understanding of the distinction between *ser* and *estar* is best achieved by considering them according to their functions, as principal verbs, auxiliary verbs or attributive (copulative) verbs. The present work focuses on the latter function, namely copulative verbs.

De Mello (1979) argues that the distinction between *ser* and *estar* as attributive verbs can be explained by taking into account their semantic content. The difference between *ser* and *estar* in their attributive functions are the features [-semantic value] for *ser* and [+semantic value] for *estar*. This means, he explains, that *ser* has no semantic value in its attributive role and *estar* has a value beyond that of its function as an attributor. He exemplifies his approach by stating that the difference between *Juan es alto* and *Juan está alto* ‘Juan is tall’ lies in the different semantic values of *es* and *está*. De
Mello (1979) states that the former example means only that *alto* is related to *Juan* and that *Juan* is related to *alto* and that no other semantic connotation is present. The verb’s only values, he states, are the expression of tense, mood, or aspect. In the latter example, he points out, the verb retains semantic value because the speaker is saying not only that *Juan* refers to *alto* and *alto* to *Juan*, but that the speaker, personally, perceives this fact, either because John was not so tall previously or because John’s tallness affects him/her in some way. De Mello (1979) explains that the words ‘previously’ and ‘affects’ provide important clues to the semantic value found in the attributive function of *estar*. This value is represented by the notion of change.

De Mello (1979) offers support to Bull’s (1942) concept of change when he states that *estar*, in contrast to the semantically empty *ser*, indicates that the idea of change is present in the mind of the speaker either because a change has taken place, or because a change could take place. He also explains that change also comes into play when a personal evaluation is made because opinion is based on a variable range of values formed by personal experience, and the value assigned could conceivably be greater or lesser than it is. The author exemplifies this by saying that to say *El regalo está bonito* ‘The gift is pretty’ to someone who has given you a gift is to admit the possibility that the gift could have been something less or more than *bonito*, but to state that *El regalo es bonito* ‘The gift is pretty’ is not to allow the possibility of the gift’s being anything other than *bonito*. In other words, he states, with *ser* the speaker feels that his/her statement is not the result of personal opinion, which is variable in nature, but rather it is an objective, impersonal statement of fact. De Mello (1979) argues that change is the manifestation of something more basic in the mind of the native speaker of Spanish in distinguishing
between *ser* and *estar* as attributive verbs, namely, the concept of temporal duration or the passage of time. He also argues that philosophical terms used previously to classify *ser* and *estar* into permanent vs. transitory, quality vs. state, and inherent vs. accidental could be misleading due to the exceptions such as *muerto* which is permanent but it is expressed in Spanish with *estar*, among the many examples found in the language.

Franco and Steinmetz (1983, 1985), similarly, propose that what is behind the distinction between *ser* and *estar* is that *ser* is more likely to be used in the context where a comparison between the individual and a group of individuals is present while *estar* is more likely to be used in the context where a comparison present is between the individual and him/herself at a different point in time. Their proposal is similar to that of De Mello (1979) in that it implies an underlying meaning for each of the copulas (i.e., change and comparison). More recently, Delbecque (1997) argues that the comparison does not necessarily have to be with the referent at another time, but that it could be between the referent and the speakers assumed beliefs or prior expectations about the referent.

Clements (2006), more recently, propose that with regard to the semantics of each verb *ser* does not carry the feature [aspect] and *estar* does. He states that in a *copula* + *adjective* construction, *estar* is used to identify the referent with the predicate complement with the feature [+locative]. Clements (2006) clarifies that the semantics of *estar* does not necessarily express the location of the referent, but rather it expresses a connection to another situation involving the referent. This connection can be assumed or expected, and it depends on the discourse situation. Clements (1988; 2006) also proposes the pragmatic feature [+nexus] which suggests that just as [+locative] of *estar*’s
core meaning implies a connection to a location, [+nexus] implies a connection to some other situation and that the connection could be assumed or expected. *Ser*, on the other hand, would not carry any of these features.

While some scholars were still trying to find a approach that might be able to explain the nuances of these two verbs, others were working on approaches that take into account the predicates the verbs appear with and the intent of the speaker when selecting either copula. Some even work with both approaches. The following sections outline the extent of such approaches.

**The adjective and its importance in the classification of ser and estar**

**Semantics**

Bull (1942) defends the thesis that the nature of the attributes as means of determining whether *ser* or *estar* would be used should be taken into account in the analysis of copula choice. He proposes a deeper look at the making of the predicate adjectives in order to be able to explain the use of one copula over the other. He also argues that in contrast with the dichotomy temporary/permanent, adjectivized past participles used with *estar* do not express the action. They express the state or condition in which the subject is found as a result of the action expressed by the participle. It represents the culmination of a process or action (i.e., the notion of change). Bull (1942) states that the subject achieves a new state and that the observer changes his concept of the state of the subject. He proposes that this dilemma can be solved by eliminating the concept of temporariness erroneously assumed, he states, to be the underlying principle behind the use of *estar* with predicate adjectives. He bases his argument in that the ‘great number of extremely obvious exceptions to the statement that adjectives are used with *estar* to indicate ‘an accidental or temporary quality or condition’ clearly point to the
false assumption underlying this rule’ (p. 434). Bull (1942) proposes that in a sentence such *El cable del ancla no estaba roto, sino cortado...* ‘The anchor’s rope was not broken, but cut off’ (Galdós in Bull, 1942) the fundamental importance in the analysis is that a change has already taken place; therefore, the underlying principle is not one of temporariness or accidentalness, but one of change. As a consequence of this proposal, Bull (1942) states, the concept of permanency for *ser* must also be abandoned. The author explains that the achievement of the new state or quality as expressed with *estar* indicates that the previous state was unstable, non-permanent; therefore, not inherent.

Bull (1942) advises that the clue to new principles is found in the type of adjective linked to *ser* and *estar* and in the semantics exhibited in their use. His approach implies dividing qualifying adjectives in two main categories. In the first one, we find adjectives that when used with ‘no’ have a positive counterpart (i.e., *no sano* ‘not healthy’ = *enfermo* ‘sick’, *limpio – sucio* ‘clean – dirty’, *caliente-frío* ‘hot – cold’, among others). He states that these adjectives possess peculiar semantic limitations because each of them measures or qualifies a portion of a range, but the positive of neither adjective can cover the whole extension of the range. In other words, there is a middle point that marks the change in meaning between each opposite or the negative of one adjective and its positive. Bull (1942) states that with these adjectives there is a change in concept. In other words, changing *feo* for *no hermoso* is a change of concept and not of subject because both adjectives cover different parts of the range.

In the second group, we find adjectives whose negative aspect cannot be expressed by the positive of another adjective. Bull (1942) uses the examples of *probable* ‘probable’, *preciso* ‘precise’, *possible* ‘possible’, and *necesario* ‘necessary’. He
claims that, with ‘probable’, if zero represents ‘not probable’ and one ‘certainty,’ there is an ascending scale of probability between zero and one because of the infinite number of fractions that exist between the two numbers. He explains that while ‘not rich’ may replace ‘poor’, ‘not probable’ cannot be replaced by the affirmative of another adjective. ‘Improbable’ must be used (i.e., another linguistic device for saying ‘not probable’). These types of adjectives, he explains, do not allow for a change in concept dealing with the same range because they cover the whole range and not just part of it. In other words, each adjective possesses its own range and make up separate semantic units. Therefore, only the positive or the negative of the same adjective are admissible not another adjective; so probable can only be substituted by not probable and not by certain. Bull (1942) added to this category adjectives that quantify nouns (i.e., limiting adjectives such as mucha/o ‘many’, poco(s)/a(s) ‘few’, etc.). Thus, for him, limiting adjectives (e.g., muchos/as ‘many’) and those negative ones that cannot be expressed by the positive of another adjective (e.g., probable ‘probable’) are used only with ser and that the other group of adjectives may be used either with ser and estar.

Moellering (1944) in an attempt to enhance the utility of the principle of subjective interpretation expressed by Andrade (1919), Morley (1925), and Bull (1942) as a basis of grammatical analysis, states: “the meaning of an adjective is a result of reciprocal influence between itself and the noun with which it is linked” (p. 598). He proposes that an adjective’s role in a noun-adjective construction is not that of modifying the noun, but of singling out some aspect of the actual object to which reference is being made and focusing attention on that aspect. He says that the joining of a noun and an adjective establishes a complex in which no real modification takes place, but that each
term makes a contribution to the total concept. The contribution is determined by the nature of the noun in the complex. Thus, when choosing between ser and estar when the attribute is predicated, the speaker indicates with ser that this attribute is one of a permanent nature in the relational constellation to which he refers with the noun employed as subject and with estar the speaker indicates that this attribute is phase-like rather than a permanent set of attributes of his present concept of the subject.

Givón (1984), moving away from the speaker’s intent towards a perception of nouns, affirms that adjectives come in antonymic pairs. These pairs can either be in a gradual-degree relation to one another (e.g., seco ‘dry’ – mojado ‘wet’) or in an absolute relation to one another (e.g., vivo ‘alive’ – muerto ‘dead’). The adjective type seco ‘dry’-mojado ‘wet’ corresponds to activity and stative verbs, according to Vendler’s (1967) classification of verbs. The type vivo ‘alive’ – muerto ‘dead’ corresponds to non-graduating states which can be said to be analogous to the notion of Vendler’s category: achievements. Givón (1984) introduces the notion of gradiency that refers to the classification of adjectives in two groups. One group is that of non-gradient adjectives or absolute adjectives (i.e., adjectives that cannot be found in ‘more or less’ constructions such as casado ‘married’ without changing the meaning of the utterance) and the other is that of gradient adjectives (i.e., adjectives that can be found in ‘more or less’ constructions such as mojado ‘wet’). For this group of adjectives and their relationship to the Spanish copula, the prediction is that gradient adjectives, because they are less absolute states, will pair more with both ser and estar because of the durative context they create. On the other hand, absolute adjectives will pair mostly with ser because of the permanent nature and the telic context they create.
As a final note in the distinction between *ser* and *estar* related to adjectives or adjective-like constructions, De Mello (1979) distinguishes the use of the *copula + past participle* as passive voice when used with *ser* and that of a resultant state when used with *estar*. The author states that this distinction may explain why certain adjectives change meaning with the use of one or the other copula.

**Individual-level predicates and Stage-level predicates**

Other approaches used to try to establish a clear-cut classification between *ser* and *estar* are related to the stage-level/individual-level hypothesis of predicate types proposed by Carlson (1977). In this approach, scholars classify *ser* and *estar* not according to the semantics of the verbs or the adjectives, but to the type of complements that accompany each verb.

Carlson (1977) defines stage-level predicates as predicates that represent stages (temporal slices) of an individual and individual-level predicates as characteristics of the individual. Individual-level predicates are atemporal and capture the notion of permanent characteristics. In this view, stage-level predicates are temporally anchored and capture the notion of temporary characteristics with more precision than prior accounts (Schmitt & Miller, 2007). Furthermore, the copula *ser* combines with the individual-level predicates (ILPs). *Individual-level* predicates express (more or less) permanent or essential properties such as *inteligente* (*intelligent*) in *José es inteligente*. *estar* combines with *stage-level* predicates (SLPs) which express temporary or accidental properties such as *cansada* (*tired-FEM*) as in *María está cansada* (Maienborn, 2005).

Kratzer, Carlson, Pelletier, Krifka, Meulen, Link, and Chierchia (1995) propose a further analysis for the individual/stage-level predicates distinction that states that predicates may differ as to the kind of eventuality underlined in the predicate. Kratzer et
al. (1995) suggest that what underlines predicates may be events, processes, or states rather than just events. The stage-level versus individual-level hypothesis encountered criticism among linguists because its similarities to the permanent/temporary distinction which guided the distinction between copulas. This approach serves to classify most of the uses of *ser* and *estar*, but it fails to move from the permanent/temporary distinction because the distinction between stages being temporary/accidental properties and individual characteristics being atemporal does not add a reasonable explanation as to why *ser* and *estar* sometimes appear with either type of predicates.

The criticism is founded in the need to take into account other aspects of languages that may help understand copula use in Spanish (Maienborn, 2005). Theoretical accounts that argue against the stage-level/individual-level hypothesis have focused their work on areas that pose more challenges to this hypothesis (i.e., variation). Most of these approaches are related to the context of *copula + adjective* and how discourse and pragmatic features help us classify the use of these two copulas in Spanish.

**Beyond the Stage-level/Individual-level hypothesis**

Maienborn (2005) states that the *ser/estar* alternation is taken as a further piece of evidence for the stage-level/individual-level hypothesis and that the distinction of SLPs and ILPs rests on a fundamental conceptual opposition that is reflected in multiple ways in the grammatical system. However, she argues, Spanish *ser/estar*, as lexical exponents of the stage-level/individual-level distinction, are promising candidates for further exploring the interpretative reflexes of the distinction.

Maienborn (2005), thus, argues against the stage-level/individual-level hypothesis and its later modifications and defends four claims as an explanation of the distribution of *ser* and *estar*. The first one is that the grammatical system is not sensitive to any
conceptual opposition like ‘temporary vs. permanent’ or ‘accidental vs. essential,’ the second is that neither ser predications nor estar predications display an underlying eventuality argument, the third one is that rather than mirroring a conceptual opposition, the ser/estar alternation is basically discourse-related: estar predications are linked to a specific discourse situation, and fourth that a discourse-based account offers a straightforward pragmatic explanation for the tendency of estar and ser predications to be interpreted in terms of the dichotomy ‘temporary vs. permanent’.

Maienborn’s (2005) argument is based on the notion that ser and estar both display the same lexical semantic properties, estar differing from ser only in presupposing a relation to a specific discourse situation. In her work, she found reasons to reject a view of the stage-level/individual-level distinction as a grammatical phenomenon with a conceptual foundation. She states that what is at the heart of the ser/estar alternation is a specificity presupposition on the topic situation. Thus, estar is the discourse-dependent variant of ser. She states that this discourse dependency is lexically triggered by estar, structurally resolved by means of the functional category of aspect, and finally, pragmatically licensed through some situation contrast found in the discourse. In other words, she states, by using estar, a speaker restricts his or her claim to a specific discourse situation, whereas by using ser the speaker makes no such restriction.

In their work on ser and estar, Clements (2006), also works with ILPs and SLPs. They propose that the distinction between ILPs and SLPs is not necessarily a semantic distinction, but rather a pragmatic one that some languages may conventionalize into a semantic distinction. He embraces Yoon’s (2001) suggestion that for Spanish it is more accurate to speak of readings given to the predicates instead of classifying them in types.
So, individual-level predicates become not a type of predicate, but a type of reading given to it and the same happens to the stage-level ones. I will adopt this terminology for the analysis conducted for this dissertation.

Clements (2006) argues that the Stage-level/Individual-level hypothesis makes the right predictions for some adjectives, but it fails to predict the correct distribution with the copulas for the ILPs *sincero* ‘sincere’, *cortés* ‘polite’, *alto* ‘tall’, *sabroso* ‘tasty’, *bonito* ‘pretty’. He proposes an alternative analysis of predicate adjectives, similar to that of Kratzer et al. (1995). This approach differentiates adjectives according to whether or not they imply some underlying process or event, and by appealing to the notion of time stability. He uses Vendler’s (1957, 1967) classification of dynamic situations\(^1\) as Activities (durative, nontelic), Accomplishments (durative, telic), and Achievements (nondurative, telic), but refers to Activities as *processes*, and to accomplishments and achievements as *events*. Clements (2006) also bases his approach on a scale proposed by Givón’s (1979) and later refined by Hopper and Thompson (1984) related to the notion of time stability. Clements (2006) explains that on the left side of this scale, parts of our world that have physical or spatial mass (e.g., mountains, stones, sky, people, etc.) are seen as prototypically more time-stable and are encoded as nouns. On the other side of the scale, parts of our world that are more transitory (e.g., the activity of singing or the punctual event of finding an object, etc.) are considered prototypically less time-stable and are generally encoded as verbs. Clements (2006) explains that between these two extremes we find the area of states, coded commonly as adjectives like *sick* or as stative verbs such as *lie, stand, know*.

---

\(^1\) Vendler (1967) classifies verbs or dynamic situations into these categories, but modern uses of this classification refer to the verb and its complement or what is called: the predicate).
At the heart of his approach, Clements (2006) state that the relevant factors for classification of adjectives are whether or not they have an underlying dynamic situation, and if they do whether it is an underlying process or an underlying event. He states that adjectives without an underlying event or process are adjectives such as *mortal* ‘mortal’, *inteligente* ‘intelligent’, and nationalities. Those with underlying process are *nuevo* ‘new’, *joven* ‘young’, *viejo* ‘old’ and the underlying process is *envejecer* ‘to grow old’. Adjectives with underlying event are those like *bautizado* ‘baptized’ for which the underlying event is *bautizar* ‘to be baptized’. He proposes that in terms of time stability, adjectives with no underlying process or event are relatively more time-stable than adjectives with an underlying process and that adjectives with an underlying process are more time-stable than adjectives with an underlying event (i.e., [+ time-stable] ‘no underlying process or event’ → ‘underlying process’ → ‘underlying event’ [- time-stable]). He establishes three predictions for the use of *ser* and *estar* with this classification:

1. Given that *ser* is unmarked for the feature [aspect], it is predicted to be most compatible with adjectives denoting states that have no aspectual properties in terms of an underlying process or event (i.e., adjectives with no underlying process or event).
2. Because *estar* is marked with the feature [aspect], it would be most compatible with adjectives denoting states that have an underlying event because events are telic and least time-stable.
3. Adjectives with underlying processes are compatible with both copulas or with neither.

Clements (2006) found problems with these predictions because even though adjectives like *casado* ‘married’ are predicted to appear only with *estar*, it is also found with *ser* (i.e., they had problems with variation). To solve their dilemma, Clements (2006) introduces the feature of directionality. He defines directionality as a feature of
adjectives with an underlying process or event; therefore, the antonyms mortal-inmortal ‘mortal – immortal’ do not possess directionality. On the other hand, antonymous pairs such as bajo-alto ‘short – tall’ assume growth and this growth is unidirectional from bajo to alto and not vice-versa. They are classified as unidirectional processes. Antonymous pairs such as sincero-insincero involve a bidirectional process because a person can move from sincere to insincere. Furthermore, antonymous pairs such as soltero-casado ‘single – married’ and acostado-levantado ‘in bed – out of bed’, both underlying events, can also be classified according to directionality. The former pair is said to be a unidirectional event and the latter a bidirectional event. Clements (2006) also finds events and processes that can be classified as either processes or events depending on the context in which they appear (e.g., bonita). In one context a woman can be bonita if she made herself look beautiful (i.e., an event), and another if the woman was not pretty before (i.e., early years) and now she is pretty, making it a process. The predictions of this hypothesis will be discussed in the following section with the inclusion of ‘the subject’ as a factor for classifying ser and estar.

Schmitt and Miller (2007) also discuss the reach of the individual/stage level hypothesis but through from a generative grammar point of view. They argue that the distinction between ser and estar is an aspectual distinction. Their claim, although different in the framework they work on is similar to Clements (2006) and Maienborn (2005), in that they all advocate for the assumption that there is a pragmatic dimension that needs to be considered. Schmitt and Miller (2007) state that the assumption is that ser is devoid of any semantic content and does not impose restrictions on the complements it can appear with. This assumption, they articulate, leads us to interpreting
ser as a state by default, unless aspectual operators are added and allow temporal interpretations of the whole ser + adjective predicates. The analysis they give for ser is closer to that of the copula as pure functional material. Schmitt and Miller’s (2007) approach proposes that estar contributes to the verb-phrase with a subevent of the type STATE. Since ser + predicate does not involve any reference to some specific interval, then the use of ser brings about the implication that the property holds independent of time. Within a generative approach, they assume that in order to assign an aspectual interpretation to the ser + adjective structure, this structure must be modified by an aspectual operator of some sort. The adverb or adverbial phrase becomes this operator because they propose that the only way to temporally anchor ser + predicate is to add an overt/covert adverbial or aspectual marker to the sentence (see Figure 2.2)

![Figure 2.2. Aspectual Marking of ser (Schmitt and Miller, 2007)](image-url)

As a result of their assumptions, they put forward two predictions. First, they propose that because estar predicates are always interpreted as STATES, ser is flexible in terms of its event type properties and that it can easily be shifted into an inchoative reading or even a temporary reading by the addition of overt or covert adverbs. Overt adverbs are adverbs or adverbial phrases that modify the utterance and covert are the ones that can be rescued from the discourse but that are not uttered (e.g., in this picture, now).
Furthermore, *ser* has no subevent property (i.e., it is not aspectually marked), depending on the context, it can be interpreted as a (a) Permanent state, (b) temporary state, or (c) an ACTIVITY (ACT BE readings). They emphasize the importance of this interpretation because it allows *ser* + predicate to describe non-inherent/non-permanent properties. In many contexts, both *ser* and *estar* may be used to describe the same situation truthfully (Schmitt & Miller, 2007). The authors propose that it is the lack of adverbs and aspectual modifiers in the sentence that triggers the implicature of permanent/inherent properties. Second, because *estar* is a state (i.e., it is aspectually marked), it triggers the implicature of temporariness associated with its use.

Schmitt and Miller’s (2007) approach on the features of *ser* and *estar* also takes into account the making of the adjectives. They propose that adjectives may have STATE-like properties but this state-like property is temporally linked only if it combines with an aspectual copula, i.e., a copula that has a subevent type, in this case, *estar*. They state that when adjectives combine with *estar* the predication is interpreted by the speaker as holding at a certain time and that the task is to find the relevant interval in which the property holds and that some permanent property can be described with *estar*, if the speaker chooses to describe it as holding at all times (e.g., Juan está calvo ‘Juan is bald’).

Schmitt and Miller’s (2007) proposal is similar to that of De Mello (1979), Clements (2006) in that neither author attaches any aspectual features to *ser* but they do for *estar* (e.g., [+semantic value], [+locative]. Schmitt and Miller (2007), thus, state that while *ser* has no inherent aspectual properties, *estar* is aspectually marked as a STATE. Their approach differs from that of De Mello (1979) in that they state that the temporary versus permanent distinction and/or the stage-level versus individual-level characteristics
of these predicates are not part of the meaning of these two verbs but rather part of the implicatures associated with the copula choice (i.e., they are related to the discourse/pragmatics realm).

Up to this point, we have seen that the discussion around *ser* and *estar* and their distribution has dealt with different properties. First, scholars were, and to some extent still are, interested in the semantics of each verb. Second, the focus shifted from the semantics of the verb to the type of predicate that accompanies them. Third, focus was put on the semantics of the adjective and its role in the copula selection. Fourth, discussion has moved out of the syntactic/semantic arena to the discourse/pragmatic one.

The subject

So far, we have seen that scholars debate the factors related to the verb and its predicate complement that help describe the distribution of *ser* and *estar*. Few of these scholars have also addressed how the subject complement of the verb affects this distribution. In this subsection, I present an outline of what has been discussed regarding the role of the subject in the *copula + adjective* construction.

Bolinger (1944) argues that the nature of the subject is an important feature on the choice between *ser* and *estar*. He puts forward three types of subjects: *infinite*, *infinitesimal*, and *evolutionary*. Infinite subjects consist of metaphysical entities such as *Dios* ‘God’, *amor* ‘love’, *fortuna* ‘fortune’, *santidad* ‘sanctity’, and other concepts when taken abstractly. This type of subject calls mainly for the use of *ser*, and for *estar* only when they refer to an everyday experience. The second category is composed of events of very limited duration as *vuelo* (flight), *verbal nouns*, and *noun clauses*. *ser* is used to indicate a first impression or a normal average concept ‘Juan es alto’ said when you first meet ‘Juan’. In this case we can say that the first impression is the only impression we get
because of the limited duration of the event. The last category is composed of facts of daily experience. They have duration and a life history. These entities call for the use of ser or estar depending on whether the attribute is regarded as normal or as a change from the norm. In sum, Bolinger’s (1944) approach states that a subject conceived as infinite or infinitesimal will call for ser before a predicate adjective and a subject conceived as evolutionary will call for the use of ser or estar, depending on whether the attribute is regarded as normal or as a change from the norm.

Clements (2006) introduced a simpler distinction for the classification of grammatical subjects that helps predict the copula that will be used. According to his theory, subjects can be classified as two types. The first one is first-order-entities (i.e., people, animals, plants, and things) and the second is second-order-entities (i.e., states, processes, and events, more likely to be located in time and to take place rather than exist). Clements’ (2006) approach states that ser is used more with first-order-entities because they have more permanent characteristics whereas estar is used more with second-order-entities because they have less permanent characteristics. Within Clements’ (2006) classification, one can find a relationship with Bolinger’s (1944) categories because infinite subjects are contained within first-order-entities, and the rest are contained in the second-order-entities.

Furthermore, Clements (2006) proposes that a combination of this notion, the notion of SL-reading and IL-reading (Yoon, 2001), and the notion of directionality of underlying events and processes in the adjective will help us to predict the choice of copula in Spanish (See Table 2.3 for a summary of the predictions).
**Table 2.3**

Copula-adjective combination according to adjective type and reading for both animate and inanimate subjects (adapted from Clements, 2006)

<table>
<thead>
<tr>
<th>Copula</th>
<th>Adjective type</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ser</em></td>
<td>adjective with no underlying dynamic situation (animate: mortal 'mortal'</td>
<td>Individual-level</td>
</tr>
<tr>
<td></td>
<td>inanimate: infinito 'infinite')</td>
<td></td>
</tr>
<tr>
<td><em>ser</em> (estar)</td>
<td>bidirectional process-adjectives (animate: sincero 'sincere'</td>
<td>Individual-level with <em>ser</em></td>
</tr>
<tr>
<td></td>
<td>inanimate: no such adjective with inanimate subject referents</td>
<td>Stage-level with <em>ser</em> and <em>estar</em></td>
</tr>
<tr>
<td><em>ser</em>/<em>estar</em></td>
<td>unidirectional process-adjectives (animate: alto 'tall' inanimate: viejo 'old')</td>
<td>Individual-level with <em>ser</em></td>
</tr>
<tr>
<td></td>
<td>Stage-level with <em>ser</em> and <em>estar</em></td>
<td></td>
</tr>
<tr>
<td><em>ser</em>/<em>estar</em></td>
<td>bidirectional event/process-adjectives (animate: bonito 'beautiful, pretty'</td>
<td>Individual-level with <em>ser</em></td>
</tr>
<tr>
<td></td>
<td>inanimate: bonito 'beautiful, pretty')</td>
<td>Stage-level with <em>ser</em> and <em>estar</em></td>
</tr>
<tr>
<td><em>estar</em> (ser)</td>
<td>unidirectional event-adjectives (animate: casado 'married'</td>
<td>Stage-level with <em>estar</em></td>
</tr>
<tr>
<td></td>
<td>inanimate: roto 'broken')</td>
<td>Individual-level with <em>ser</em> (only animate subject referents)</td>
</tr>
<tr>
<td><em>estar</em></td>
<td>bidirectional event-adjectives (animate: cansado 'tired'</td>
<td>Stage-level with <em>estar</em></td>
</tr>
<tr>
<td></td>
<td>inanimate: abierto-cerrado 'open-closed')</td>
<td></td>
</tr>
</tbody>
</table>

Variation also posed a challenge to Clements’ (2006) approach because he found that while event-adjectives and adjectives without an assumed underlying dynamic situation seem to behave in the expected manner, the adjective *extenso* ‘extensive’, with
an implied process of knowledge accumulation appears, not with *estar*, but with *ser*.

They also found that second-order entity subjects behave somewhat differently than first-order entity subjects and conclude that they may be differently constrained. Even though Clements’ (2006) approach found explanation to most of the cases in which variation is found, variation has proven to be problematic for theoretically-based approaches to copula distribution in their approach and those discussed above.

**The adverb**

An element of a sentence whose role in the distribution of copulas in Spanish has not been fully explored is the adverb or adverbial phrase (i.e., adjuncts). This element has helped explain phenomena related to the acquisition of tense and aspect by second language learners (Bardovi-Harlig, 1992; Bardovi-Harlig & Reynolds, 1995) and may very well help us explain the distribution of *ser* and *estar*. At the present time, only one study related to first language acquisition has taken this element into account. Schmitt and Miller (2007) propose, as mentioned above, that because *estar* predicates are always interpreted as STATES, *ser* is flexible in terms of its event type properties and that it can easily be shifted into an inchoative reading or even a temporary reading by the addition of overt or covert adverbs. Overt adverbs being adverbs or adverbial phrases that modify the utterance and covert the ones that can be rescued from the discourse but that are not uttered (e.g., in this picture, now). The reach of their prediction is discussed in the following section.

**Previous Work on the variation of the structure *ser* and *estar* + adjective**

Two fields have seen a growth of models to explain the use of copula choice from different points of view than that of theoretical linguistics and prescriptive grammar. These fields are Sociolinguistics and Second Language Acquisition of Spanish. In this
section, we focus on sociolinguistic studies that provide the basis for my own study regarding copula use in bilingual speakers of Limon. The participants of this study are not second language learners of Spanish because they learn both languages from birth. The fact that they are fully bilingual separates the present study from those of second language acquisition of Spanish. Therefore, I will not discuss here second language studies. For a full account of copula choice in Spanish SLA the reader is directed to Geeslin (2002a, 2002b; 2003, among others).

Previous analyses have proposed that variation between *ser* and *estar* is a change in progress (Silva-Corvalán, 1986) accelerated by the language contact situation between English and Spanish in bilingual communities in the US (i.e., Los Angeles). More recent studies have examined *ser/estar* variation in other dialects of Spanish in contact with other languages (Geeslin & Guijarro-Fuentes, 2007, 2008; Guijarro-Fuentes & Geeslin, 2006) and monolingual varieties of Spanish (i.e., Venezuelan Spanish) where linguistic and extra-linguistic evidence points out that *ser/estar* variation is stable (Geeslin & Diaz-Campos, 2005b). All three will be outlined in this section which is divided into studies on Spanish in contact with English, studies in monolingual Spanish, and Spanish in contact with other languages.

**Spanish in Contact with English**

Silva-Corvalán (1986) represents the first and most important study to explore the extension of *estar*. She studied the speech of 27 bilinguals of different generations and degrees of Spanish language proficiency. Data were collected through recordings of conversations between the researcher and the participants. There were three groups of participants, those who were born in Mexico and migrated to the U.S.A. after the age of eleven (Group I); those who were born in the U.S.A. or migrated to the U.S.A. before the
age of six (Group II); and those born in the U.S.A. whose parents were of Mexican ancestry who fit the definition of the second group (Group III). She states that the choice between *ser* and *estar* does not depend only on syntactic or lexical constraints, but that the extended discourse and shared knowledge among the interlocutors must be considered.

In her study, Silva-Corvalán (1986), with the use of cross-tabulations, examined data using variables such as adjective class which divides adjectives according to their semantic categories, class and individual *frame of reference* (Falk, 1979). This variable distinguishes whether the adjective used to describe a subject classifies this subject as part of a class of subjects with the same characteristics or not. Another variable used was *susceptibility to change* which categorizes attributes of the subject as susceptible to change or not. *Animacy* also was included. This variable divides subjects into animate beings and inanimate ones, and semantic transparency which classify adjectives depending on how transparent their meaning is to the speaker. The last category contains four subcategories of different modality which refers to instances in which there is no semantic difference in the adjective. One of these subcategories is *apparent synonymy* which states that there are no differences the proposition with either copula, choice not allowed in standard language, and clear difference.

She found that adjectives differ with respect to the innovation. Adjectives of *size*, *physical appearance*, *age*, and *evaluation* favor innovative uses and that the innovative use of the first three classes correlated with *animacy* of the subject; that high frequency of *estar* with certain attributes may be promoting its extension to less frequent *class frame* contexts in which the same attributes are used. She also found that contexts where the
contrast between *ser* and *estar* is transparent with a group of adjectives that have different meanings associated with the choice of copula were more resistant to the extension of *estar*. Furthermore, she found that in contexts where the copula choice is not associated with a semantic difference in the adjective, but a subtle difference in the verb, favors the extension of *estar*; that apparent synonymy favors the extension of *estar*; and that the innovative uses of *estar* increase as the level of proficiency in Spanish decreases.

She concludes that the innovation present in her study represents part of an evolutionary trend in Spanish and other Romance languages and that language contact accelerates this trend. She states that the condition of reduced access or lack of access to formal varieties of the language must be met in order for this diffusion to happen and that internally motivated changes which involve generalizations across languages are also accelerated in a situation of extended contact. With regard to the copula choice, Silva-Corvalán (1986) suggests that ‘the extension of *estar* in progressives [*estar* + present participle] and its frequent association with *be* in those constructions [*be* + v-ing] may favor the rapid diffusion of *estar* in the context of predicate adjectives, where Spanish has evidenced a slow process of change independent of any language-specific influence’ (p. 604). She suggests that there is no noticeable movement toward a steady functional specialization; but that the continuous renovation of Spanish due to the arrival of other immigrants keeps Los Angeles Spanish from changing entirely. She concludes by saying that the Spanish copula opposition with attributes is lost to a large extent among speakers in her Groups II and III, but that these speakers are unlikely to pass on Spanish to their descendants. This led her to assume that the system described in her study which
maintains a limited amount of meaningful variation will persist as a defining feature of English-dominant bilinguals.

Ortíz-López (2000) studied another variety of Spanish in contact with English. His study focused on the Spanish of Puerto Ricans. He followed Silva-Corvalán’s (1986) approach taking into account semantic transparency, age, education and adjective class as factors. He studied the responses to a 47-linguistic-context questionnaire by 122 participants and 20 interviews by another group of participants. The participants were 41 males and 81 females for the questionnaire group and eleven males and nine females for the interview groups. They were sixteen balanced bilinguals, sixty-seven intermediate and thirty-nine participants with poor domain of English. He used cross-tabulations to analyze his data. His results showed that innovative use accounted for nineteen percent of this sample on the questionnaire and twelve percent on the interviews. He studied the innovative form of *estar* with one linguistic variable: adjective class. He found that innovative uses are more common with adjectives of status, followed by adjectives of size and evaluation for the questionnaire. For the interview data, he reported that innovation was found more with adjectives of age followed by adjectives of size. He reports that men use more *estar* than women in the questionnaire, but that differences were not as evident. In the interview, he found that women are more resistant to change (i.e., they use *estar* less). One important finding of this study is that Puerto Rican bilinguals use *ser* more than those that were not total bilinguals. He states that in the case of Puerto Ricans, contact with English is not a variable that leads to change, which is different from the bilinguals from Los Angeles. They are also different from those of speakers from Houston. Ortíz-López (2000) explains this difference by stating that
Puerto Rican Spanish has been acquired as an L1 and its acquisition has been supported by formal education in Spanish. The support from formal education in Spanish, he states, neutralizes or stops the force that English could have as a conditional factor. He proposes that semantic transparency aids the extension of *estar* especially with adjectives that imply a constant change. These adjectives are adjectives of age, size, physical appearance, evaluation, among others. He adds that semantic transparency is heightened if we incorporate syntactic elements such as temporal adverbs like *ahora* ‘now’, which add a semantic function of temporality when choosing the verb. He concludes that this change is caused by internal linguistic constraints such as the type of adjective and by syntactic and pragmatic factors such as contact and age.

Similar to the studies mentioned above, Salazar (2007) studied the extension of *estar* from a innovative-prescriptive approach. Unlike previous studies of this type, Salazar used variable rule analysis in her data. Her data came from the New Mexico Colorado Spanish Survey (Bills & Vigil, 1999). She chose twenty participants (8 male and 12 female). Their ages range between 50 and 70+. Seven of her participants had eight years or less of formal education and the rest eight or more. Her dependent variable was the use of *ser* or *estar*. Her independent variables include conservative or innovative use, verb tense, adjective type, presence or absence of an intensifier, presence or absence of a time adverb, and code-switching after the copula (i.e., the adjective was said in English). She found that contact varieties show higher percentage of innovative use of *estar* when compared to other studies in monolingual Spanish (i.e., the three studies presented above). She also found that adjective type, especially adjectives of age, was the strongest predictor of *estar*. The second strongest predictor was presence or absence of a time
adverbial followed by code-switching, and the presence or absence of an intensifier. Her variable verb tense was not selected as significant. Among the social factors that showed to be significant predictors of *estar* were education. Gender was not significant and she eliminated age as a variable because she found incoherent results and states that more data would be needed to evaluate this factor group. She concludes that despite being categorized as an isolated and archaic variety of Spanish, New Mexico Spanish exhibits innovative use of the copular verb *estar* in combination with adjectives and participles and that adjective type and the presence of an adverbial are the factors most likely to cause a speaker to choose *estar*.

Salazar’s (2007) findings are important in three ways. First they validate the hypothesis that contact with English accelerates the rate of change and adds to the body of knowledge already established for varieties in contact with English. Second, she provides empirical evidence to Ortíz-López’ (2000) and de Jonge’ (1993), discussed below, hypotheses that the presence of a time adverbial would contribute to the presence of *estar*. However, these findings contradict Schmitt and Miller’s (2007) hypothesis that adverbs would be needed with *ser* and not with *estar*. Because of this discrepancy between a theoretical account and an empirical study, more studies are needed to further test these two hypotheses. Third, it is the first to use variable rule analysis in a variety of Spanish in contact with English.

**Monolingual Spanish**

Gutiérrez (1992, 1994a; 2003) found, in the Spanish of Michoacán, Mexico, patterns similar to those found by Silva-Corvalán (1986). He studies the speech of twenty-six monolingual Mexicans from the same social class as those studied by Silva-Corvalán (1986). He used percentages of use as a point of comparison. He found that the
frequency of innovation was lower in monolingual Spanish, but that the fact that innovation was present supports the presence of a change. He found similar comparative patterns to those of Silva-Corvalán (1986). This finding contrasts to that of bilingual communities he compared them to in that these adjectives are found with the innovative form alongside moral characteristic, social status, perception and color. He concludes that although innovative *estar* is still competing with *ser* in several areas, evidence from the Michoacán case points out to a very significant movement towards innovation (p.175) and that this innovation is not adding a new meaning. He also concludes that evidence from the bilingual communities of Los Angeles and Houston give further evidence in favor of the extension of *estar*. He found this extension in higher frequencies which allowed him to state that language contact accelerates changes that have been originated in monolingual environments in agreement with Silva-Corvalán’s (1986) findings.

De Jonge (1993) compared the two varieties of monolingual Spanish, Caracas Spanish and Mexican Spanish, in order to test his hypothesis that language change does not happen as a continuous phenomenon but rather in different stages including periods of complete stop. His data come from two sources. The Mexican data comes from the Mexico Culto Corpus from the sixties and a questionnaire applied in the eighties to 60 Mexican students of the Universidad Nacional Autónoma de México (UNAM). The Caracas data comes for the Caracas Culto corpus collected at the beginning of the seventies and a wave of data collected in the eighties. He uses chi-square and odds ratios as his statistical analyses. He frames his study in the grammaticalization of structures and language change. He identifies two adverbials that helped him test his hypothesis. These adverbs are *ya* ‘already’ and *cuando* ‘when’. He states that the identification of
estar with ya/cuando should be taken as the first step on the way to grammaticalization. He found that a frequency of one cuando to eight ya’s is not sufficient for a contamination (i.e., extension) of the contexts that contained cuando with estar in Mexican Spanish which brings the process of change to a temporary stop. For Caracas Spanish, it was a frequency of one cuanto to three ya’s which does not allow the use of estar to expand any further until a new candidate in the contexts with estar comes around for extension to happen. He concludes that his data show that the extension of estar depends on the circumstances within which the new form appears. He states that if an element occurs often enough in particular contexts, extension would happen so that the “train of language change can move ahead to the next station” (de Jonge, 1993). Finally, he states, if no new element at hand can bridge the gap to the next station, extension stops.

De Jonge’s (1993) hypothesis of language change is important because it may explain why different varieties of Spanish tend to yield somewhat different results (see Guijarro-Fuentes and Geeslin, 2008 below), but does not contradict the fact that change progresses slowly in monolingual varieties and faster in bilingual varieties. His inclusion of the adverb as a possible explanation for the extension of estar contradicts Schmitt and Miller (2007) hypothesis that estar has a temporal feature related to time because he found that adjectives like “ya” and “cuando” are found with high frequency with estar and Schmitt and Miller’s (2007) theoretical prediction is that estar would not need the presence of such adverbs.

Malaver (2001) studied copula choice with expressions of age in the Spanish of Caracas. Her purpose was to provide a semantic-pragmatic characterization of the
functions of age expressions in this variety of Spanish (i.e., Caracas Spanish). She also analyzed the relationships among linguistic and social factors in *copula + expression* of age constructions. She studied the speech of 208 speakers from two different corpora (1977 and 1987) collected in Venezuela. Her sample included an even number of men and women (n=104 each). For the 1977 corpus, she used two social classes and for the 1987 corpus she used five social classes. Participants’ ages range between fourteen and more than sixty years old. She analyzed a total of 527 instances of *ser* and *estar*. Her dependent variables were whether the expression of age was neutral or prototypical. Her independent variables included the adverb with which it was found (e.g., *ya* ‘already’ with *estar* = prototypical and *cuando* ‘when’ with *ser* = neutral), and whether a determiner such as *uno* ‘one’, *la* ‘the-feminine’ was present in the expression of age (prototypical with *ser*), and *mas* ‘more’ or *muy* ‘very’ (prototypical with *estar*). She analyzed her data using descriptive statistics and Chi-squared tests. She found that *estar* use increased in the ten-year period between corpora in expression of age. She concluded that the presence of prototypical and neutral expressions of age with *estar* is evidence of a functional difference in the conception of expressions of age and that this phenomenon is typical of Spanish of the Americas because it has only been documented in Caracas and Mexico. She also concluded that it is a stable change that was already present in the Spanish of Caracas in the 19th century.

Cortés-Torres (2004) adds to this body of knowledge her findings in a new variety of monolingual Spanish: Cuernavaca Spanish. She studied the extension of *estar* through a variable rule analysis of the data collected from thirty-six participants. She collected data of two sorts. The first one was oral data and the second one was a questionnaire. She
coded her variables as innovative or not innovative use. Her variables included the linguistic variable adjective class and she included adjectives of *age, size, physical appearance, evaluation and personal characteristics*. She also included social variables such as *level of education, age, gender and style*. The variable style came from both oral data and questionnaire. She found that adjectives that favor *estar* are adjectives of physical appearance, age and evaluation whereas adjectives that disfavor it are adjectives of size, personal characteristics. Social variables that favor *estar* are education and style. Participants with lower level of education favor the use of *estar* whereas the ones with higher levels of education disfavor it. Conversation style disfavors the use of *estar* while questionnaire favors it. She concludes that there is an innovative use of *estar* in the Spanish of Cuernavaca.

Cortés-Torres’ (2004) finding regarding style seems problematic because she reports that all contexts in the questionnaire “were framed as a framework of class” (p. 790) (Falk, 1979) which makes the results from this instrument be highly favorable of *estar* while the conversation was a less biased towards one of the copulas. However, her findings regarding the presence of an ongoing change in Cuernavaca, Mexico is an important contribution to the study of copula choice in Spanish because it adds a new variety of monolingual Spanish.

Díaz-Campos and Geeslin (in press) applied the multi-feature model proposed in Geeslin’s line of work (2002a, 2002b) to study the copula choice in a variety of Spanish, said to have no contact with English: Caracas, Venezuela. They used the Estudio Sociolinguistico de Caracas2 (1987) [financed by the Consejo de Desarrollo Científico y

---

2 Sociolinguistic Study of Caracas
Humanístico de la Universidad Central de Venezuela\textsuperscript{3} which includes half-hour interviews, conducted in 1987 and 1988 of 160 speakers born and raised in Caracas, with parents also from Caracas. Each interview was divided evenly among four age groups, both genders and five socioeconomic groups. The authors studied data from four age groups: 14-29, 30-45, 46-60, 61 and above, three social classes (low, middle and upper); and they were evenly divided between men and women (2 participants per cell). They concluded that the multivariate analysis they proposed showed that the factors, both linguistic and non-linguistic, included in their analysis are relevant in describing variation in copula use in Caracas Spanish. These variables were \textit{predicate type}, \textit{susceptibility to change}, \textit{experience with the referent}, \textit{resultant state}, \textit{adjective class}, \textit{copulas allowed}, \textit{age}, and \textit{social class}. They also found that older speakers tended to favor the use of \textit{estar}, a pattern of behavior that is an indication of the stability of this phenomenon in Caracas Spanish.

Geeslin and Diaz-Campos (2005b) investigated whether the extension of \textit{estar} was a change in progress or a stable change. They researched how copula use differed between older and younger speakers by studying the frequency of use of \textit{estar} and its predictors. They also studied how frequency of occurrence of the predictors of \textit{estar} would vary across age groups. In a re-analysis of the 2004 data, the authors found that older generations tend to favor the use of \textit{estar} and there are no prominent differences between female and male speakers. In their predictor analysis by age group they found that \textit{predicate type}, \textit{resultant state}, \textit{adjective class} and \textit{copulas allowed} were common predictors for all age groups. However, there were different predictors that were significant, but for specific age groups. These predictors were \textit{susceptibility to change} for

\textsuperscript{3} Scientific and Humanistic Development Council of the Universidad Central of Venezuela
age group 14-29; experience with the referent for groups 14-29, 46-60, and 61 and above; and socioeconomic class for group 46-60. They concluded that a sociolinguistic interview does not elicit the same types of contexts from all speakers, and that three of the variables show a greater proportion of the categories that favor estar for older speakers. Their results on age and gender seem to indicate that the extension of estar is a stable phenomenon, but they also found that types of contexts were not equally represented among age groups and that a comparison across age groups must also take these differences into account. Their results revealed that discourse factors such as predicate type, resultant state, adjective class, and copulas allowed are more important predictors than age and gender.

Díaz-Campos and Geeslin (2005a) studied the extension of estar with adjectives and its relationship to language contact. They examined the individual categories of the adjective class variable to determine their relationship with the use of estar. Their intention was to relate adjective classes to the innovative use found in previous work. They stated that the classes that are the best predictors of use of estar may not actually represent innovation. They were guided by questions such as what categories of the adjective class variable show the most frequent use of estar; what categories of the adjective class variable show innovation; and how their results relate to previous studies on adjectives and copula choice. In a subsequent analysis of the data, they found that adjective classes with highest use of estar were mental state, size, physical appearance, color, status, and age. The distribution of estar with each adjective class across the variable copulas allowed showed more tokens of estar in the both allowed category than were required. Their results revealed uses of estar in contexts formerly reserved for ser
for all age groups. A high frequency of use of *estar* is especially noticed with adjectives of *age*, *size*, and *status*. Furthermore, when both copulas are allowed speakers tended to favor *estar*, evidence that innovative uses are emerging. They concluded that the monolingual Caracas Spanish shows similarities to US/Mexican Spanish in the innovative use of *estar* with adjectives of *age*, *size*, *physical description*, *evaluation*, *color* for certain age groups, for adjectives of *status* and *size* for all age groups, and *age* and *description of personality* for the older speakers.

Díaz-Campos and Geeslin’s (in press), Cortés-Torres (2004), and Díaz-Campos & Geeslin’s (2005; 2005a) works are important because they are the first ones to use a multi-feature analysis that includes analyses beyond the cross-tabulations used by other authors. They introduce a new way of analyzing variation by using a set of variables that come from different theoretical accounts as well as from previous studies on copula choice that include syntactic, pragmatic and discursive variables.

**Spanish in contact with other languages**

In order to add more information to the extension of *estar*, Guijarro-Fuentes and Geeslin (2006) studied the responses of 37 bilinguals from Galicia, Spain to a contextualized questionnaire where they were required to select the appropriate copula according to the discourse context that was provided to them. The authors searched for the linguistic features that best determine copula for Galician speakers of Spanish, to what degree bilingualism determines copula choice, and compared how Galician speakers of Spanish compare to other Spanish-speaking population, more specifically those studied by Guijarro-Fuentes and Geeslin (2003). To achieve their goal, the authors collected social data via a background questionnaire and linguistic data via a written contextualized preference task that contained 28 items, each of which is preceded by a
paragraph-length context. Three data analysis procedures were conducted. The first one was a set of Chi-square tests used to determine which linguistic variables, in isolation, were strong predictors of copula choice. The results showed that all linguistic variables proved to be strong predictors of copula choice when taken in isolation. The second analysis was a regression analysis conducted to test which of these variables were strong predictors of copula choice when embedded in a regression model. The results of this regression showed that predicate type, frame of reference, susceptibility to change, adjective class, and copulas allowed were strong predictors of copula choice. Social variables were analyzed in the same fashion as the linguistic variables. However, when in isolation, none of the social variables turned out to be strong predictors of copula choice. When embedded in the regression model, only two social variables were strong predictors of copula choice: native language and age. The authors conclude that Galician bilinguals respond to the same linguistic contextual features in using estar as monolinguals. They also found that few of the social variables related to bilingualism were good predictors of copula choice and suggested that more research should be done to account for their influence in copula choice.

A large-scale study on Spanish in contact with four languages of the Iberian Peninsula was conducted by Geeslin and Guijarro-Fuentes (2008). They set out to address estar selection in situations of language contact in Spain in order to extend the body of research beyond contact situation between Spanish and English. The first of its kind, Geeslin and Guijarro-Fuentes (2008) provided a large-scale, multi-group comparison of copula selection in Spanish in contact with Catalan, Galician, Valencia, and Basque. They were guided by the questions of how and with what frequency copulas
are selected in each of the five language regions. They also explored to what degree
speakers within a single group respond unanimously in their selection of estar. They
continue the line of research by asking what the linguistic predictors of estar were and
whether there were differences between the monolingual variety and each of the bilingual
varieties. Their data comes from five different groups. The speaker groups were eighty-
three Spanish monolinguals, seventy-three Galician-Spanish, sixty-six Valencian-
Spanish, one-hundred and forty one Catalan-Spanish, and seventeen Basque-Spanish
speakers. All the participants in the study filled out a background questionnaire that
elicited information about social characteristics, their language learning histories, and
current patterns of language use. They completed a contextualized questionnaire in
Spanish which contained 28 items. Each item was preceded by a context and asked the
participants to choose between ser, estar or both copulas. All contexts were part of a
single story.

The authors used three types of statistical analysis on their data. The first one was
a Chi-square to determine whether the frequency of estar selection was different for each
bilingual group as compared to the monolingual group. Second, a cross-tabulation of
response type for each given item was conducted for each population in order to identify
those responses where members of a particular group are unanimous and those where
variation within a single group occurs. Third, a binary-logistic regression analysis was
done for each group in order to determine which factors are the best predictors of estar
selection when considered in a single model. Their dependent variables were ser, estar or
both. Independent variables included predicate type, frame of reference, experience with
the referent, animacy, susceptibility to change, adjective class and copulas allowed.
Extralinguistic variables included *residence, education, gender, second language(s), age, first language(s), mother’s first language, father’s first language, childhood first language, language of education, normal language, work language, home language, and social language*. They excluded the variable *profession* because of the high number of academic-related participants in their sample. They divided age into ten-year increment groups in order to capture a wide range of age groups. The authors found that speakers of Catalan and Valencian had lower rates of use of *estar* while speakers of Basque and Galician showed higher rates of use than the monolinguals. Each bilingual group was statistically significantly different than the monolinguals and all groups showed variation in copula selection. After conducting the regression analysis, they found that the variables *predicate type, adjective class and copulas allowed* were strong predictors of *estar* for all groups. *Susceptibility to change* was a strong predictor for Catalans, Valencians and monolinguals. *Dependence on experience* was significant for Catalans, Galicians and monolinguals. For monolinguals variables that were strong predictors also included *animacy* which makes monolinguals the group with more strong predictors of *estar*. They found that bilingualism does not have the uniform effect of increasing the frequency with which *estar* is selected. The found that *animacy* seems to have irregular effects across group and that *dependence on experience* appears to be linked to the size of the participant group where larger groups show a statistically significant effect and smaller group do not. They conclude that no single variable or combination of variables can distinguish variable contexts from categorical ones. In other words, they stated it means that additional factors, or simply individual variation, are likely relevant in further predicting copula selection (p. 376).
Geeslin and Guijarro-Fuentes (2006; 2008) are very important for the present study; first, because they introduce a new contact situation. Second, because they use a more complex type of analysis. Third, because they are the first large-scale study that compares five varieties of Spanish. Furthermore, the point made by Geeslin and Guijarro-Fuentes (2008) regarding group size and their effect on the results of the analyses conducted represent one of the reasons why the proper sample size calculation is needed during the design stages, a point I discuss in detail in Chapter 3.

Methodological Issues

As mentioned in Chapter 1, one objective underlying the structure of this literature review was to search for key components used to calculate sample size. The other was to select factors that would help me construct a model for the prediction of estar in contexts of copula + adjective which will be discussed in Chapter 4. The importance of looking for these key factors comes from the differences found in the results of the previous studies. Sample size differences would yield different results like those found by Geeslin and Guijarro-Fuentes (2008) between language groups and across studies of copula choice. These differences could lead to mixed interpretations and contradicting hypotheses regarding the processes of change. Tagliamonte’s (2006) note of caution summarizes this issue. In order to make her point in favor of an alternative to statistical significance when looking for differences between factors, she states that

“in some situations, statistical significance does not provide the best evidence for interpreting results. A data set with a larger number of tokens will tend to detect more factors to be statistically significant than one with fewer tokens. This means you should not compare the results of parallel analyses in two or more communities on the basis of significance alone.”(p. 237)
Tagliamonte’s (2006) assertion of the problem related to differences in sample size and its relationship to the results serves as the justification for the following critique. The body of knowledge regarding copula choice in Spanish has grown in the last twenty-five years. A representative sample of these types of studies was presented in this chapter. For comparison purposes and as a visual aid to the extent of the sample size problem, I created Table 2.4 to summarize each of the empirical studies presented above. This table shows the author and year, variety of Spanish studied, sample size used for both speakers and tokens, statistical analyses used, and factors that were significant.

From Table 2.4 we can observe that samples sizes have not been consistent across studies or within studies. This issue poses problems to the generalizability of each study’s results to the population at large and renders comparison of results across varieties of Spanish somewhat difficult. As we can see, different studies with different sample sizes produce different significant results. Even those studies comparing different populations with different sample sizes produce different significant results.

To solve this type of problem in the field of Psychology, Wilkinson and the APA Task Force (1999) created a set of guidelines for the reporting of statistical findings. One of these guidelines refers to the importance of reporting sample size and the calculations used to reach those numbers. All but one study in Table 2.4 report their sample sizes, but none reports the statistical rationale for choosing the number of participants they chose. This issue poses another problem for the comparison of the results of different studies carried out for different populations. If the calculations of sample size are not reported, the power of the study is unknown and the generalizability of their results is hindered which makes the comparison difficult.
<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Spanish</th>
<th>Variety of Spanish</th>
<th>Other Language</th>
<th>Sample size Speakers</th>
<th>Sample Size Tokens</th>
<th>Type of data</th>
<th>Statistical Analyses</th>
<th>Significant Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silva-Corvalán (1986)</td>
<td>Contact</td>
<td>Los Angeles</td>
<td>English</td>
<td>27</td>
<td>623</td>
<td>oral</td>
<td>Cross-tabulations</td>
<td>Susceptibility to change, adjective class, semantic transparency, and animacy</td>
</tr>
<tr>
<td>Ortiz-Lopez (2009)</td>
<td>Contact</td>
<td>Puerto Rico</td>
<td>English</td>
<td>322</td>
<td>952</td>
<td>questionnaire Cross-tabulations</td>
<td>semantic transparency, Adjective class, education</td>
<td></td>
</tr>
<tr>
<td>Salazar (2007)</td>
<td>contact</td>
<td>New Mexico</td>
<td>English</td>
<td>20</td>
<td>45</td>
<td>oral</td>
<td>Varbitul</td>
<td>Adjective class, time adverbial, code-switching, education, and intensifiers</td>
</tr>
<tr>
<td>De Jonghe (1993)</td>
<td>Monolingual</td>
<td>Caracas 70's, 89's</td>
<td>Mexico</td>
<td>not reported</td>
<td>309</td>
<td>oral</td>
<td>Cross-tabulations</td>
<td>Adjectives of age, adverbial and cuando</td>
</tr>
<tr>
<td>Malave (2001)</td>
<td>Monolingual</td>
<td>Caracas 77</td>
<td>48</td>
<td>67</td>
<td>Interview Percentages of use and Chi-Square</td>
<td>expressions of age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cortés-Torres (2004)</td>
<td>Monolingual</td>
<td>Caracas 87</td>
<td>Cuernavaca</td>
<td>160</td>
<td>460</td>
<td>Interview Varbitul</td>
<td>Adjectives, education, style</td>
<td></td>
</tr>
<tr>
<td>Díaz-Campos and Geeslin (2004)</td>
<td>Monolingual</td>
<td>Caracas (all)</td>
<td>48</td>
<td>782</td>
<td>oral</td>
<td>Chi-square, logistic regression</td>
<td>predicate type, susceptibility to change, experience with the referent, resultant state, adjective class, copulas allowed, age and social class</td>
<td></td>
</tr>
<tr>
<td>Geeslin and Díaz-Campos (2005a)</td>
<td>Monolingual</td>
<td>Caracas (Young)</td>
<td>24</td>
<td>956</td>
<td>Logistic regression</td>
<td>Adjective class: mental state, size, physical appearance, color, status, and age. Copulas allowed more often.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geeslin and Díaz-Campos (2005b)</td>
<td>Monolingual</td>
<td>Caracas (all)</td>
<td>48</td>
<td>826</td>
<td>Logistic regression</td>
<td>predicate type, resultant state, adjective class, copulas allowed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geeslin and Díaz-Campos (2005b)</td>
<td>Monolingual</td>
<td>Caracas (14-29)</td>
<td>12</td>
<td>826</td>
<td>Logistic regression</td>
<td>Susceptibility to change, experience with the referent.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geeslin and Díaz-Campos (2005b)</td>
<td>Monolingual</td>
<td>Caracas (46-60)</td>
<td>12</td>
<td>826</td>
<td>Logistic regression</td>
<td>Experience with the referent, socioeconomic class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geeslin and Guijarro-Fuentes (2006)</td>
<td>Contact</td>
<td>Caracas (61+)</td>
<td>Spain</td>
<td>37</td>
<td>Logistic regression</td>
<td>Experience with the referent, predicate type, frame of reference, susceptibility to change, and copulas allowed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geeslin and Guijarro-Fuentes (2006)</td>
<td>Contact</td>
<td>Spain</td>
<td>83</td>
<td>7296</td>
<td>questionnaire Chi-square, logistic regression</td>
<td>Predicate type. Adjective class, copulas allowed, susceptibility to change, animacy</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>341</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>426</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The last methodological issue that can be pulled out of the design of these studies is the statistical analysis used in each study. Most studies seek to explain the relationship of *estar* and certain linguistic and extra-linguistic factors while others try to predict the use of *estar* in the context of *copula + adjective*. Most of the studies presented above use cross-tabulations and Chi-square which provide the level of association of one variable with the variable under study. Several use a type of logistic regression known as variable rule analysis, which is designed to calculate the results in a form most useful in variationist studies (Tagliamonte, 2006). Some use a binominal step-wise regression found in SPSS. All three of these analyses are appropriate for the questions that each study sets out to answer. However, the use of different methods of statistical analysis also makes the comparison between studies hard to conduct because their results come from methods that have completely different assumptions. A full discussion of these issues is presented in Chapter 3.

**Conclusion**

The review of relevant previous literature regarding the *copula* construction points to six important issues. First, variation found with respect to copula choice has posted numerous challenges to proponents of theoretical accounts which has made the classification of this variation as a sociolinguistic variable difficult. These accounts, however, are the theoretical foundations from which one must start in search of an explanation for the distribution of *ser* and *estar* because they have been demonstrated to be very helpful in the construction of a statistical model for the explanation of such phenomenon. Second, in order to better understand copula choice in Spanish, we need to use a design such as those employed by Silva-Corvalán (1994), Guitiérrez (1994a), Geeslin and Díaz-Campos (2005b), Díaz-Campos and Geeslin (in press), in
sociolinguistics and Schmitt and Miller (2007) in first language acquisition that takes into
account different linguistic and social features. These studies have shed light on the fact
that there is syntactic variation found in copula choice across varieties of Spanish, both
monolingual and in contact with other languages. Third, in any new endeavor in search of
an explanation for the copula choice in Spanish, all different factors should come from
well-founded theoretical accounts and experiments as attested by the design of the studies
mentioned above. Fourth, for the study of copula choice in Spanish, syntactic, semantic,
and pragmatic approaches must be used per claims of the different scholars dealing with
this particular phenomenon. However, some methodological issues (e.g., sample size)
need to be addressed to reach the desired levels of generalizations needed to claim a
complete or near complete theory. These issues are addressed in Chapter 3. Fifth, the fact
that monolingual and contact Spanish show similarities with regard to copula choice
(Díaz-Campos & Geeslin, in press; Guijarro-Fuentes & Geeslin, 2006; Silva-Corvalán,
1986, 1994) serves as justification to study the Spanish of Limón for it is a contact
situation that has not been studied before. There is a lack of evidence related to copula
choice in contexts in which Spanish is an official language and English is not. Finally,
another justification for conducting the present work is the debate of whether it is a
change in progress (Morley, 1925; Silva-Corvalán, 1994) or a stable change (Geeslin &
Díaz-Campos, 2005b). The study of Limonese Spanish may provide further evidence for
either of these hypotheses because it is a variety of contact Spanish where both a
monolingual and a bilingual variety of Spanish dwell in the same geographical region.
All six issues serve as the basis for undertaking the present study and each is addressed in
the following chapters. In order to answer the questions derived from this line of work
presented in Chapter 3 and restated in Chapter 4, all methodological issues related to the methodology of this type of study needed to be addressed in the early stages of its design.

In brief, there have been many approaches to explain the differences in use of ser and estar and as Moellering (1944) put it “the ‘last word’ on ser and estar will never be printed.” My intention is not to do so, but to enrich this already rich area of inquiry. In order to model what seems to be a structure with multiple faces, additional studies, as stated above, are needed. The goal of this dissertation is to provide the field with an extra building block for our understanding of copula choice in the context of copula + adjective in Spanish, to provide the field of sociolinguistics with a tool for the calculation of proper sample size for future research based on representativeness and generalizability of results, and to propose a modification to the variable rule analysis that takes into account the structure of the data. In Chapter 3, I deal with four methodological issues that were evident in the empirical studies on copula choice in Spanish and that were addressed in early stages of this work.
Chapter 3

Methodological Issues Addressed

Introduction

From the review of the previous literature on copula choice in Spanish, three research questions are asked for the present work:

1. What linguistic variables predict the use of *ser* and *estar* + adjective in the Spanish spoken by Costa Ricans in Limón?
2. What social variables predict the use of *ser* and *estar* + adjective in the Spanish spoken by Costa Ricans in Limón?
3. Can the pattern of variation in the use of *ser* and *estar* be considered a change in progress or a stable change?

In order to answer these questions, several steps were taken for the creation of a sound design that fulfills the suggestions provided by the American Psychological Association for the generalizability of research findings. The first step was to address several key methodological issues that had not been addressed in previous empirical studies on copula choice in Spanish. Thus, after reviewing previous empirical work done on this topic; four methodological issues were of relevance to the design of the present work. I addressed them during the early stages of this work for reasons that are presented throughout this chapter. These issues are: (1) the importance of a proper power analysis during the design stages, or the probability of correctly rejecting the null hypothesis when it is false in the population, (2) the analysis of the structure of the data to select the proper statistical analysis to be used, (3) polychotomization, or categorization, of continuous variables and its effect on the power of the study, and (4) the determination of sample size to achieve an adequate level of power. Each of these issues was selected to be addressed here due to their importance in reaching generalizable results, from an
empirical point of view, following the latest standards in the fields of psychology and statistics. After a careful analysis of the review of the previous literature and the lack of reports on power and sample size calculations to support the generalizations derived from each study, I present a thorough discussion of the theoretical foundations of my design decisions. This chapter is divided into four sections. Each of the sections deals with one of the issues mentioned above and presents the theoretical background and its relationship to the field of applied linguistics in general that led me to make decisions that structured the methodology addressed in Chapter 4.

Applied Linguistic studies and sample size planning

In the fields of linguistics, sociolinguistics, and second language acquisition power of a study or the derivation of sample size calculations are rarely mentioned in published reports. What are reported are the efforts to justify the use of convenience samples or samples that have been collected in order to fill a predetermined grid. Silva-Corvalán (2001) points out that “[la] cuestión del número óptimo de individuos que asegure la validez y representatividad de la muestra es aún un problema no resuelto en sociolinguistica” (finding the optimum number of subjects that guarantees validity and representativeness to the sample is still an unresolved problem in sociolinguistics). However, sample size planning, as the main goal of this chapter, is an issue that needs to be addressed in the design stages of any investigation. In addition, the procedure used to determine it should be reported in empirical research reports (Wilkinson & the APA Task Force, 1999) per the American Psychological Association suggestions. The field of linguistics or applied linguistics should not be an exception to this sound practice. It is evident, from the review of the previous literature for this work and claims in research

---

4 My translation
design textbooks (Silva-Corvalán, 2001) that this issue has in reality not been resolved; what seems to be the norm is that sample size is taken as a matter of numbers or incorrect assumptions regarding statistical tests. In other words, all other things being equal, the more subjects/tokens we have the better the study will be; or, even worse, a matter of a random set number per cell. This misconception derives from the belief that to increase power we need to increase the sample size without regard to other theoretical and empirical implications or the fact that for certain statistical tests (e.g. Chi-square) an expected number per cell is desired.

Silva-Corvalán (2001) discusses the issue and states that traditionally in sociolinguistics a number of 25 subjects for every 100,000 speakers or .025 of the population under study based on Labov (1967) has been the norm. She argues that, instead, the answer to sample size should be determined by theoretical issues such as the nature of the problem and the resources that the sociolinguist has to carry out his/her investigation. This claim is in accord with what most researchers in the fields of psychology and statistics propose and what I implement in this investigation. She continues her discussion stating that it seems that five speakers per cell is considered adequate if each cell represents a homogeneous social subgroup. This practice derives from Lavandera’s (1975) claims that if the objective of the researcher is not to discover the general frequency of a linguistic phenomenon in a community, but to establish correlations between linguistic variation and extra-linguistic characteristics of the population, the sample need not include a proportional number of individuals with regard to each social group in the population. She states that an equal number of subjects in each category is sufficient. More recent developments in the field of statistics for the social
sciences (i.e., hierarchical models) have demonstrated that this procedure hinders the power of a study and may yield spurious results (Maxwell & Delaney, 1993), an issue to be discussed momentarily. The drawbacks of such practice and its implications in the generalizability of the results are addressed throughout this chapter.

I will take Silva-Corvalán’s (2001) suggestion that sample size planning should be based on theoretical and practical issues and follow what is recommended in the field of statistics for the social sciences for the determination of a proper sample size. My interest in this section is not to criticize the practice in applied linguistics, but to introduce the concern that the practice followed in applied linguistics may be producing underpowered studies which, in turn, may be yielding generalizations that lead to incorrect conclusions due to their weakness with regard to design. We can also be getting more subjects that is necessary to address the problem. Or, from a budgetary point of view, we are spending more money than is really necessary in the collection of data beyond what is really needed to explain our phenomena due to a poor design. The following sections address the procedure to determine the proper sample size for the present sociolinguistic study following the latest advances in the field of statistics for the social sciences.

**Importance of a proper power analysis during the design stages**

In order to understand power analysis, we first have to understand the importance it has and the consequences of producing underpowered studies. In the following sections I address the definition of power, the importance of power analysis and how to approach it appropriately in any field of inquiry, and consequences of not doing so in an attempt to be able to transfer this practice into the field of applied linguistics.
Power

Power is defined as the probability of correctly rejecting the null hypothesis and it is the complement of the probability of Type II errors⁵ (Kelley, Maxwell, & Rausch, 2003) and what Cohen (1969, 1988, 1992a, 1992b) called the probability that a statistical test will yield statistically significant results. This definition has been taken up by several authors in their endeavor to promote the adoption of scientifically sound practices within the behavioral and social sciences (Beck, 1994; Kraemer & Thiemann, 1987; Leon, 2004; Lipsey, 1990; Onwuegbuzie & Leech, 2004; Rossi, 1990; Schmitz, Cherny, & Fulker, 1998; Sedormeier & Gigerenzer, 1989; Zodpey, 2004) and by the American Psychological Association’s guidelines for publication (Wilkinson and the APA Task Force, 1999).

Importance of power analysis

Murphy & Myors (1998) propose that an understanding of statistical power first requires understanding the ideas that underlie statistical hypothesis testing. They state that one of the most basic ideas in statistical analysis is that results obtained in a sample do not necessarily reflect the state of affairs in the population from which that sample was drawn (p. 2). They also point out that statistical tests should be taken as decision aids and that these tests can help us draw conclusions about whether the findings of a particular study represent real population effects, or whether they are the results of random sampling errors. Lipsey (1990) suggests that the results of hypothesis testing (i.e., whether we reject or fail to reject the null hypothesis) is not a matter of probability. He states that either they differ or they do not (i.e., not one more significant than the other), and that if we knew the population parameters we could determine the differences between them directly without any statistical test.

---

⁵ Type I and Type II errors are defined below
In order to understand Lipsey’s (1990) suggestion in the context of applied linguistics studies, I will address the issue of relying too heavily on the notions of reject or prove the null hypothesis (i.e., reliance in the significance level) and what seems to be an overgeneralization of statistical packages results rather than a proper design issue (i.e., the notion that 0.05 is worse than 0.01 and even worse than 0.001).

Cohen (1988) approaches power, after conducting a post-hoc power analysis, as a response to the growing number of studies in psychology done with the null hypothesis as a framework that yielded very low power. He states that we need to acknowledge the probabilistic character of statistical inference and to set aside the reliance on the words reject and prove. Cohen (1988) suggests that this might be done by requiring that every researcher set certain appropriate probability standards, during the design phase of the study, for research results to provide a basis for rejection of the null hypothesis and for the proof of the existence of the phenomenon under study (i.e., .05 or .01 level of significance). He states that results from a random sample drawn from a population can only approximate the characteristics of the population and that, even if the null hypothesis is really true, a given sample result is not expected to represent the population exactly. Even though, Cohen’s (1988) concern was related to the field of psychology and the social sciences, we can extend it to the field of Linguistics and its subfields. A brief look at the most recent publications is sufficient to see that an over-reliance on the words reject and prove and generalizations without proper power analyses are rooted in our field. It is evident when generalizations are made based on the notion of significant difference and p-values of .05 or less without any report of how each researcher established the significance level during the design stages of the study and when in a
single study different $p$-values are reported. None of the studies presented in Chapter 2 reported power levels, sample size calculation, or significance-levels in the design stages.

Rossi (1990) declares that due to the fact that the aim of behavioral research is to discover important relations between variables, a researcher must consider power as natural and an important part of the planning and interpretation stages of research. Cohen (1992b) adds that the importance of power analysis arises from the fact that most empirical research in the social and behavioral sciences is done by formulating and testing null hypotheses which researchers wish to reject as a means of establishing findings about the phenomena studied. Algina & Olejnik (2000) point out that the benefits of power analysis are noticeable at two levels. One is that sample-size selection is made on a rational basis rather than using guidelines that may or may not apply in a particular situation and the other is that the researcher specifies the size of the effect that is substantively interesting or expected. Cohen (1992b) states that the exact definition of the nature of the phenomenon’s existence depends on the details of how it is manifested and statistically tested. He also says that it is convenient to conceive the significance criterion as embodying both the probability of falsely rejecting the null hypothesis and the definition of the existence of the phenomenon when necessary. In other words, the researcher has to be aware of how conservative s/he wants to be when taking chances of rejecting the null hypothesis during the design process. This level of conservatism has been established as a level of .05 being not very conservative and a level of .01 or below as being very conservative. By setting this level a priori, the researcher demonstrates how founded her/his conception of the phenomenon is and how much risk s/he wants to take in search of answers to her/his questions.
Another approach to power analysis and hypothesis testing is that of design sensitivity. Within this framework, Murphy & Myors (1998) state that the power of a statistical test is a function of its sensitivity, the size of the effect in the population, and the standards or criteria to test statistical hypotheses. To accomplish this, Murphy & Myors (1998) suggest that the simplest method to increase the sensitivity of a study is to increase its sample size which makes the estimates become more precise. They also suggest that different treatments have different effects and that power increases as the effect size increases; therefore, it is easier to reject the null hypothesis when the significance level is set to .05 than when it is set to .01. Power increases as the significance level becomes less stringent. For design sensitivity, power is highest when all three conditions (i.e., sensitivity, effect size, and criteria to test statistical hypotheses) are met and sample size becomes the most important element to determine it.

Low power studies can have severe consequences at different levels of generalizability. Murphy & Myors (1998) point out that experimental designs that have low power which, in turn, render lack of replicability cannot accomplish their central purpose of determining the effects of treatment or, in this case, prediction of association. In studies where sample size is smaller than recommended, researchers may not have enough power to reliably detect the effects of even the most substantial treatments. These studies are more likely to make Type II errors, or the probability of falsely rejecting the null hypothesis when it is true, because their sample size is not large enough, have small effect sizes, or have very strict significance levels. Determining power and the significance test allow for the probability of rejecting the null hypothesis if the same
analysis is run over and over again which can only be done when the researchers’ theory
is true (Kraemer & Thiemann, 1987).

Large samples can also bring negative consequences to the power of a study. Murphy & Myors (1998) point out that with a sufficiently large sample size almost any result would be statistically different from zero; therefore, almost any null hypothesis tested, provided the null hypothesis is false, would be rejected. They also state that large sample sizes make statistical tests highly sensitive with the same results.

In sum, power represents a vital piece of information about a statistical test applied to research questions (Cohen, 1988). The call for using this piece of information when designing a study is well founded and deserves the attention of any researcher who wants to advance any field of science because their contribution will be more explanatory of the phenomena under study; therefore, the aim of this dissertation is to bring this issue to the attention of my fellow linguists who work with empirical data. Therefore, Labov’s (1967) call for a sample size of .025 of the population, Lavandera’s (1975) justification for an equally distributed sample, and Silva-Corvalan’s (2001) call for the simple collection of more data need to be revised in order to reach a design that properly addresses both the general frequency of a linguistic phenomenon in a community and the correlations between linguistic variation and extra-linguistic characteristics of the population. This design is given by a proper research design based on the power of the study (i.e., power analysis).

Power Analysis

The power of a statistical test depends upon three parameters: the significance criterion, sample size, and the effect size. In other words, the power of a statistical test depends on the degree to which the phenomenon exists (Cohen, 1988). Kelley &
Maxwell (2008) define power as a function of four things: the size of the effect (i.e., unstandardized), the model error variance (i.e., standardized), the Type I error rate, and sample size. Each of these components will be fully described in the following subsections. Their applicability to the field of applied linguistics, specifically to this study, is presented in the last section of this Chapter.

**The statistical test or Statistical Analysis**

Lipsey (1990) states that because the determination of statistical significance (i.e. alpha) and estimation of the probability of error in the statistical conclusion are made within the framework of a particular statistical test, the test itself is one of the factors determining statistical power. Thus, different statistical tests have different statistical power when they are applied to the same data. In other words, each statistical analysis we apply to our data has its own power. Therefore, in order to carry out a proper research design the analysis to be carried out and the power of such test should be set during the design stages and not when the data have been collected. Because the statistical test is closely related to the significance criterion, the discussion of both components of power analysis is presented together.

**Significance Criterion (Alpha level)**

Lipsey (1990) points out that the level set for the significance criterion or alpha (i.e., the $p$ level) influences the likelihood of statistical significance. A larger alpha makes it easier to reach significance. The likelihood of significance, when the null hypothesis ($H_0$) is true, increases with each increase in alpha. Cohen (1992b) states that the probability of mistakenly rejecting the $H_0$ or Alpha ($\alpha$) is a researcher decision and it represents the maximum risk s/he is prepared to take of making this error. Because it has become a convention, the risk is set at .05 by default but other levels can be used (e.g.,
.10, .010, .0010, etc.). However, these levels are, as pointed out by Lipsey (1990), representative of the risk one is willing to take and not how significant a test is, and as I mentioned above they are either different or not, but not more or less different/significant as sometimes presented in applied linguistics reports.

Sprinthall (2003) and Lipsey (1990) recognize four possible outcomes with regard to testing $H_0$. The first one is that the researcher rejects it when it should have been rejected, because the null hypothesis is false in the population (i.e., there is a significant difference) which makes this the correct decision. The second is that the researcher fail to rejects it when it should not have been rejected because in the population it is true (i.e., there is no significant difference) which also makes it a correct decision. The third one is when the researcher rejects it when it should have failed to reject it because in the population it is true (i.e., claims a difference, but in reality there is none) which makes it an incorrect decision. This is a Type I error. The last one is when the researcher accepts it when it should have been rejected because in the population it is false (i.e., claims there is no difference, but in reality there is) which makes it an incorrect decision. This is a Type II error. The outcomes can be summarized as shown in Figure 3.1.
As we can see in Figure 3.1, the probability for a Type I error is called alpha (α) and the probability of a Type II error is called beta (β). In simple terms, Type I errors translate into a conclusion derived from our sample that a phenomenon does not exist when it really exists in the population, but it was not detected due to sampling error. Type II errors are when, in the same fashion, a phenomenon is said to exist when in reality it does not, but it was detected, again, due to sampling error. Within the design of this dissertation a Type I error would be that due to a sampling error or design, I would reject the influence of, for example, “gradiency” as a factor in the prediction of the use of copula, but in reality, this variable does help predict the use of copula. Type II errors, on the other hand, would be that because of an error in the design or determination of sample

---

6 Adapted from Lipsey (1990)
size, I would accept the influence of this variable in the prediction of copula use, but in reality, this variable does not influence the use at all.

Both Type I and Type II errors should remind us that hypothesis testing does not offer a fool-proof, full-service, absolute guarantee (Lipsey, 1990); therefore, we need to be cautious in the interpretation of the results. Lipsey (1990) also points out that neither rejecting nor accepting $H_0$ represents the last step, but that the statistical decision within the context of the Type I and II errors does provide a better chance than the mere flip of a coin. Thus, a design based on these assumptions provides a more sound approach to the generalization of findings and the analyses of what is and what is not present in the data.

**Effect Size**

Vaske, Glinert & Morgan (2002) define effect size (ES) as the strength of the relationship between the independent variable (i.e., $X$) and the dependent variable (i.e., $Y$). They also add, citing work by Rosenthal (1994), that the computations for ES have been divided into two major types often referred to as the $d$ family and the $r$ family of indices. They explain that the former group are expressed in standard deviation units and computed by finding the difference between the means of two groups of interest and dividing by the appropriate standard deviation (i.e., $d = \frac{\mu_1 - \mu_2}{\sigma}$) and the latter is expressed as a correlation coefficient (i.e., $r$). In the second method the effect sizes are always between -1 and +1. The former type represents the effect of instruction in an experiment that seeks to determine whether a specific teaching method is effective. The latter is taken as the degree of association that a set of variables have with a particular phenomenon. Thus, in this dissertation the effect size will be the degree of association between the independent variables and the use of copular verbs in Spanish.
Lipsey (1990) states that if $H_0$ is false, there is some real difference between the experimental conditions. This difference or effect can be measured and its size has an important influence on the likelihood of attaining statistical significance. He points out that the larger the effect, the more probable its statistical significance and the greater the statistical power holding everything else constant. Leon (2004) stipulates that even though power analysis is simple for a design comparing means for two independent groups, it is more complex for more complicated designs like the one attempted in this dissertation. This assertion is due to the fact that finding the effect size in the social sciences is, in Cohen’s (1992b) words, “the most difficult aspect of power analysis” (p. 99). Cohen (1992b) blames this difficulty on the low level of consciousness about magnitudes in the social sciences. He also blames the work of psychological science, specifically the work by Fisherian null hypothesis testing, because it has had the unfortunate effect of emphasizing the magnitudes of probability values from significance tests rather than the magnitudes of the phenomena under study. This has resulted in, what was discussed above, an over-reliance on the words reject or prove, instead of the interpretation of the real effect of the variables in the phenomenon we are trying to explain.

Murphy & Myors (1998) point out that effect sizes in the social and behavioral sciences tend to be small or moderate which results in the need for large sample sizes. They state that if the effect of a treatment is large enough to be detected in small samples, there is little reason to test for that effect. The authors also state that not all effects are large enough to be detected in small samples and the tendency is not to abandon the traditional criteria for statistical significance accepted in their fields of inquiry. They
argue that because of this tendency, effect sizes and decision criteria tend to be similar across a wide range of studies while samples sizes vary considerably impacting directly the levels of power. An example of this variation in sample size and length of the effect sizes under study is Otheguy, Zentella, and Livert (2007). In their study on overt and covert pronouns in Spanish in New York, they collected data from more than three hundred speakers producing a database of approximately sixty-five thousand instances of their dependent variable, yet they found no significant results with respect to the effects of variables such as age, gender, social class, etc. Other studies, on the other hand, find significant results of the effect of similar variables with relatively small sample sizes. In the line of studies of copula choice in Spanish, most significance levels are reported to be either .05 or .01 while effect sizes are not reported. Even though these measurements tend to be very similar and the effect size should be similar, theoretically, sample sizes are not. For example Diaz-Campos and Geeslin (in press) report using 48 subjects (4 age groups, 2 participants per cell, male and female) while Guijarro-Fuentes and Geeslin (2006) reported using 37 participants and both found significant results. See also Table 2.4 in Chapter 2 for further examples.

Cohen (1992b) states that positing what population effect size to use comes from the investigator’s knowledge of the field. In other words, it comes from sample effect sizes found in previous investigations with similar variables, the results of pilot studies, and his or her educated intuition. He proposes a nomenclature to classify effects sizes: small, medium, and large. His intention was to state that a medium effect size is an effect size that can be apparent to the naked eye of a careful observer, a small effect size can be noticeably smaller but not trivial, and a large effect size is the same distance above the
medium size as the small is below it. These indices vary according to the statistical test conducted and their meaning is also established according to the phenomenon under study. Table 3.1 summarizes Cohen’s (1992b) suggestions for classification of effect sizes.

Because the aim of this dissertation is not to discuss the interpretation of effect sizes, but rather to draw attention to the need to know the meaning of it within a specific context, I direct the reader to the work of Cohen (1992b) and Grissom and Kim (2005) for a full discussion on effect sizes and their interpretation.

Table 3.1

Effect Size Indices according to Statistical Test

<table>
<thead>
<tr>
<th>Test</th>
<th>Effect Size Indices</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
<td>Medium</td>
</tr>
<tr>
<td>Null Hypothesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population parameter = Ø</td>
<td>.10</td>
<td>.30</td>
</tr>
<tr>
<td>Departure of a population proportion (P) from .50</td>
<td>.10</td>
<td>.25</td>
</tr>
<tr>
<td>ANOVA $\mu_1 = \mu_2$</td>
<td>.20</td>
<td>.50</td>
</tr>
<tr>
<td>$t$-test</td>
<td>.20</td>
<td>.50</td>
</tr>
</tbody>
</table>
In sum, effect size is the discrepancy between the null hypothesis and the alternative hypothesis of interest. Every statistical test has its own effect size index, which is zero when the null hypothesis is true. This difference is absolute for a two-sided test and is either positive or negative for a one-sided test. Each standardized effect size is a pure and scale-free value that measures the discrepancy between the null hypothesis and the alternative hypothesis or population parameter (Cohen, 1992b).

**Timing for power analysis**

Regarding the time when this analysis or calculation should be done, Zodpey (2004) points out that the issue of power can be addressed at two stages of the study. The first one is before the data are collected and using appropriate approaches and information on effect size and significance level in order to determine the sample size needed to achieve these parameters. The second is after the data are collected and during the interpretation of the results as a *post-hoc* analysis of power. There is still discussion on the pros and cons of *post-hoc* power analyses; therefore, most scholars in the field of statistics for the social sciences recommend it be done a priori.

**Structure of the data and selection of proper statistical analyses**

In any field of inquiry, it is recommended to use appropriate statistical methods to analyze data. The choice of method is closely related to the research questions and to the structure of the data to be analyzed. In other words, research questions should guide the statistical analyses performed on the data, and they should be appropriate to the structure of the data in order to arrive at the correct answer. This is not to say that in linguistics these procedures are not followed, but to call attention to the reporting of such procedures and calculations to allow for the replicability of studies. What follows is a description of
the data-structure collected for the present work followed by the description of the proper statistical analyses selected for it.

In this section, I introduce the background of analyses followed in the field of sociolinguistics and potential drawbacks related to their use to analyze the data gathered for the present work and relevant statistical advances for the analyses needed to approach such data. It is an introduction to the statistical procedures used in the field of sociolinguistics and how they have helped analyze language data in an appropriate way in previous years. However, these procedures are not free from scrutiny and what I present here is a call for the revision of such analyses with the desire of improving the statistical methods employed in the study of language variation and the type of data linguists deal with, especially in corpus data.

Variable Rule Analysis (VARBRUL)

Cedergren and Sankoff (1974) showed that appearance of a feature or subcategory tends to affect rule frequency in a probabilistically uniform way in all the environments containing that feature (Tagliamonte, 2006). They reach this conclusion from the need to explain language variation and the findings in Labov (1969) where he found that the choice process in language operates regularly across a wide range of internal and external contexts and from his statement that the form of the grammar itself is a set of quantitative relations in language. With Labov’s (1969) and Cedergen and Sankoff’s (1974) findings, the use of inferential methods for the analysis of language phenomena became more frequent, especially in the field of sociolinguistics. The use of these methods did not come as easily as might have been expected, due to the fact that statistical procedures such as ANOVA were not appropriate to the analysis of language data (Tagliamonte,
Other procedures had to be used, so the currently used method called Variable Rule Analysis or VARBRUL was designed.

Tagliamonte (2006) describes how VARBRUL started as formal expressions compatible with the apparatus of formal language theory of the time (i.e., Generative grammar rules). She also describes that the reference to variation as a rule was related to variation being systematic and not just a mere formalism. Tagliamonte (2006) explains how variable rules are not rules in their roots, but rather, citing Sankoff (1988b, p. 984), “the probabilistic modeling and the statistical treatment of discrete choices and their conditioning” (p. 131).

Tagliamonte (2006) also explains that Variable rule analysis, in order to account for the extreme distributional imbalances that include contrasting full vs. near-empty cells in corpus data, uses the likelihood criterion (i.e., the comparison of observed values of the dependent variable to predicted values obtained from models with and without the independent variable in question). This criterion measure allows the researcher to see how likely it is that a particular set of data has been generated by the model which has the given values for the independent variables and their effects. Thus, different sets of factor effects will have different likelihood measures for the same set of data. This point becomes very important for the calculation of sample size within this type of analysis, a point discussed below.

As Tagliamonte (2006) points out, the estimation of logistic regression via maximum likelihood\(^7\) is not unique to linguistics. She points out that the analysis of linguistic data requires a modified variation of logistic regression, which is provided by a

---

\(^7\) finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum (Weisstein, 2009)
statistical package called GoldVarb (Sankoff, Tagliamonte, and Smith, 2005) that calculates the results in a form that serves better the goals of this particular field. Silva-Corvalán (2001) explains that VARBRUL is designed to manage low number data and that this mathematical procedure converts real frequencies of the dependent variable according to the independent variables in theoretical probabilities and assign a probabilistic index to each independent variable. This index reflects the effect of each factor on the probability that the dependent variable occurs. She points out that VARBRUL uses a multiple logistic regression and that the values range between 0 and 1. The interpretation is that higher values represent a positive correlation (usually .5 and up) and lower values (usually lower than .5) a negative correlation between the occurrence of the dependent variable and the independent one.

GoldVarb, as explained by Tagliamonte (2006), uses two types of analysis. The first one is a binomial one-step that allows the researcher to examine each of the cells at the same time to see how much each combination (i.e., each \( Y \)) differs from the expected. The second is a binomial step-up/step-down, which performs a leveled analysis in which computations are done one step at a time. This provides the researcher with three lines of evidence containing statistical significance, relative strength, and constraint ranking factors. For example, Aguilar-Sánchez’ (2007) results show that the variables “gradiency” was the second variable in importance in the prediction of \( estar \) with probabilistic weight of .398 for gradient adjectives and .668 for non-gradient ones (constraint ranking factors) with a log likelihood of -80.602 (relative strength) and a \( p \)-value of .014 (statistical significance). The analyses performed during the step-up/step-down include the calculation of how each factor included in the model helps to explain
the variance. This is done one factor at a time until the analyses reach the saturated model (i.e., a model that includes all the predictive variables). The step-down procedure performs the analysis, but in the opposite direction: starting with the saturated model and ending with the null model (i.e., a model with no predictors). In order to calculate the inclusion of each factor group, the software uses the Chi-square test to test for significant reductions in the likelihood function.

This statistical analysis has become widely used among sociolinguists; however, the structure of the data to be analyzed does not always meet the assumptions made. In Tagliamonte (2006), the explanation of the analysis only includes factor groups at the sentence or word level, but does not include variables of a different nature (i.e., non-linguistic variables). The advice is to keep the data from the individual in a different file so that the results by individual (i.e., different runs for each individual) can be compared or contrasted (Tagliamonte, 2006, p. 168). However, the practice in the field is to clump individual-level and linguistic-level variables together and then analyze the results at one level (see Aguilar-Sánchez, 2007; among others; Díaz-Campos & Geeslin, in press; Tagliamonte, 2006).

Hox (2002) explains that the practice of including these two levels together brings two distinct problems to the analysis of this type of data. The first one is statistical. Hox (2002) states that if data are aggregated, the result is that different data values from many sub-units are combined into fewer values for fewer higher-level units. The author states that the result is that much of the information is lost and the statistical analysis loses power. If the contrary happens, if the data are disaggregated, Hox (2002) states that the result is that a few data values from a small number of super-units are exaggerated into
many more values for a much larger number of sub-units. What this means is that if we count, for example, each coding of “female” or “male” in the sample of individuals producing several observations each as separate accounts of “female” and “male”, we will be generating more values of the factor “gender” than what there actually are in the population. We will be treating highly correlated observations as independent observations. Hox (2002) emphasizes that any statistical test will treat these disaggregated values as independent information from the much larger sample of sub-units. He declares that the appropriate sample size for these variables is the number of higher-level units. In my example, this will be the total number of males and females (i.e., participants) in the study and not every instance of “female/male” coding attached to the observations. Hox (2002) also points out that “using the larger number of disaggregated cases for the sample size leads to significance tests that reject the null-hypothesis far (emphasis is mine) more often than the nominal alpha level suggests. In other words: investigators come up with many ‘significant’ results that are totally spurious” (p. 3).

The second problem is, according to Hox (2002), conceptual. He explains that if the researcher is not very careful in the interpretation of the results, s/he may commit the fallacy of the wrong level, which consists of analyzing the data at one level, and formulating conclusions at another level. He continues by stating that the best-known fallacy is the ecological fallacy, which is interpreting aggregated data at the individual level (p. 3). Another fallacy that can occur, he states, is the atomistic fallacy or the fallacy of formulating inferences at a higher level based on analyses performed at a lower level. In more general terms, Hox (2002) proposes that the problem of drawing completely
erroneous conclusions may arise if grouped data, drawn from heterogeneous populations, are collapsed and analyzed as if they came from a single homogenous population. This type of problem is also known as the “Simpson’s Paradox” (Hox, 2002). In other words, if we take the data of “female” and “male” in the present investigation as representing different sources to each of our tokens, then we will be drawing erroneous conclusions regarding gender because all the tokens coming from the same speaker are highly correlated with regard to gender. Thus, a second-level analysis would help us to avoid this mistake.

Hox (2002) explains that in this type of data structure the individual observations are, in general, not completely independent. In other words, and for this dissertation, all instances of ser/estar produced by an individual are not completely independent from one another. This will result in the average correlation between variables measured on each speaker being higher than the average correlation between variables measured from different speakers. Hox (2002) continues by stating that “standard statistical tests lean heavily on the assumption of independence of the observations and that if this assumption is violated the estimates of the standard errors of conventional statistical tests are much too small which, in turn, results in many spuriously ‘significant’ results” (p. 5). Multilevel data, most of the time, presents this problem.

Thus, knowing the structures of your data to select the appropriate statistical analysis for it can solve the problems that arise from the structure of the data. For data that have a hierarchical structure such as the data collected for this dissertation, a multilevel analysis in the context of logistic regression is an appropriate method (Gelman & Hill, 2007; Hox, 2002) or for other statistical tests (Raudenbush & Bryk, 2002).
The data in the present investigation are representative of what is called a multi-level data set. It is multi-level because data come from different levels. The first level is the level of the token. Each individual interviewed for this project produced a number of instances of ser and estar + adjectives. The second level is the level of the individual him/herself (e.g. gender, social class, education level, etc.). A possible third level is any grouping that may be assigned according to the characteristics of the individual (e.g., contact language as presented in Geeslin and Guijarro-Fuentes, 2008). Thus, any sociolinguistic data set represents a multilevel set of data. Because of the dichotomous nature of the dependent variable, the only statistical analysis that can be employed is that of logistic regression. What follows is an explanation of logistic regression, its contextualization within the multi-level analysis of data, and its relationship to sample size calculations.

Statistical Analyses

In order to describe the type of procedures used for the analyses of the data for this dissertation, I present an overview of the traditional statistical procedures as well as the latest advances for the analysis of data types like mine. The first subsection is an introduction to logistic regression and its assumptions, the second is a description of the fit of the logistic regression model, the third one is an explanation of significance testing for the logistic regression model, and the fourth one is a brief explanation of the multilevel logistic regression analysis used to analyze the data collected for the present work.

Logistic Regression

Hosmer & Lemeshow (1989) state that regression methods have become an integral component of any data analysis concerned with describing the relationship
between a response variable \((Y)\) and one or more explanatory variables \((X)\). They voice that it is often the case that the outcome variable is discrete, taking on two or more possible variables. Kleinbaum, Klein, and Pryor (2002) define logistic regression as “a mathematical modeling approach that can be used to describe the relationship of several \(X\)’s to a **dichotomous** dependent variable” (p. 5). Hosmer & Lameshow (1989) explain that logistic regression has become the standard method of analysis in this situation and that before beginning a study of logistic regression, it is important to understand that the goal of an analysis using this method is the same as that of any model building technique used in statistics. In other words, they state that the goal is to find the best fitting and most parsimonious, yet substantively reasonable model to describe the relationship between a dependent variable and a set of independent variables, which are often called covariates.

Furthermore, Hosmer & Lemeshow (1989), Kleinbaum, Klein, & Pryor (2002), and Vittinghoff (2005) state that the difference between linear regression and logistic regression is that in the latter the dependent variable (DV) is binary or dichotomous. All authors agree that this difference is reflected both in the choice of a parametric model and in the assumption they make. One of the differences is accounted for in the methods employed in an analysis using logistic regression, which follows the same general principles employed in the linear regression. In addition, Hosmer & Lameshow (1989) state that when the researcher looks at the scatterplots as the way to begin exploring the relationship of the dependent variable and the independent variables (IV), in the logistic regression the plot depicts the dichotomous nature of the dependent variable and does not provide a clear picture of the nature of the relationship. Thus, in order to remove some
variation and maintain the structure of the relationship, the researcher needs to create intervals for the independent variables and compute the mean of the dependent variable within each group. This way, two important differences may emerge. The first one is the nature of the relationship between the DV and the IVs. In any regression problem the key quantity is the mean value of the DV given the value of the IVs. This quantity is called the conditional mean and is expressed as $\hat{Y}_X$ where $\hat{Y}$ denotes the DV and the subscript $X$ denotes the value of the IV. This expression is read as the expected value of $Y$, given the value of $X$. It is also represented as $E[Y|X = x]$, where $x$ is a realization of the variable $X$. With dichotomous data the conditional mean must be greater than or equal to zero and less than or equal to 1 or $[0 \leq \hat{Y}_X \leq 1]$ for a realistic model. A logistic distribution is used to provide a model for $\hat{Y}_X$ in the case when $Y$ is dichotomous.

The quantity $\pi(X) = \hat{Y}_X$ is used to represent the conditional mean of $Y$ given $X$ as

$$prob[Y = 1|X = x] = \pi(x)$$

Thus, the specific form of the logistic regression model is as follows:

$$g(X) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_px_p$$

or

$$\frac{e^{\beta_0 + \sum_{k=1}^{K} \beta_k x_k}}{1 + e^{\beta_0 + \sum_{k=1}^{K} \beta_k x_k}} = \pi(x)$$

which after a transformation central to the study of logistic regression called the logit transformation for $k$ variables becomes:

$$g(x) = \beta_0 + \sum_{k=1}^{K} \beta_k X_k$$
This transformation is important because \( g(X) \) has many of the desirable properties of a linear regression model. The logit, \( g(X) \) is linear in its parameters, may be continuous, and may range from \(-\infty\) to \(+\infty\), depending on the range of \( X \). Christensen (1997) clarifies that the logit transformation takes a number \( p \) between 0 and 1 and transforms it to \( \log \left( \frac{p}{1-p} \right) \), that the logistic transformation takes a number \( x \) on the real line and transforms it to \( \frac{e^x}{1+e^x} \), and that they are the inverse of each other. Christensen (1997) affirms that doing an analysis of the data requires both of these transformations.

Hosmer and Lameshow (1989) state that the other important difference concerns the conditional distribution of the DV. In the linear regression model, they explain, one assumes that an observation of the DV may be expressed as \( Y = \hat{Y}_X + \varepsilon \) where \( \varepsilon \) is the error and expresses an observation’s deviation from the conditional mean. The most common assumption is that \( \varepsilon \) follows a normal distribution with mean of zero and some variance that is constant across levels of the IV. Thus, the conditional distribution of the DV given \( X \) is normal with mean of \( \hat{Y}_X \), and a variance that is constant.

However, in the case of the dichotomous DV this is not the case. When the DV is dichotomous, one may express the value of the DV given \( X \) as \( Y = \pi(X) + \varepsilon \). According to Hosmer & Lemeshow (1989), if a nominal scaled variable has \( k \) possible values, then \( k - 1 \) design variables will be needed. Hosmer & Lemeshow (1989) state that the quantity \( \varepsilon \) may assume one of two possible values. If \( Y=1 \) then \( \varepsilon = 1 - \pi(X) \) with probability \( \pi(X) \), and if \( Y=0 \) then \( \varepsilon = -\pi(X) \) with probability \( 1 - \pi(X) \). Therefore, \( \varepsilon \) has a distribution with mean zero and variance equal to \( \pi(X)[1 - \pi(X)] \). This means
that the conditional distribution of the DV follows a binomial distribution with probability given by the conditional mean, $\pi(x)$.

Hosmer and Lameshow (1989) summarize the logistic regression analysis as:

The conditional mean of the regression equation is formulated to be bounded between zero and 1. 
The binomial, not the normal, distribution describes the distribution of the errors and will be the statistical distribution upon which the analysis is based. 
The principles that guide an analysis using linear regression will also guide the logistic regression analysis.

The Fit of the Model

Kleinbaum, Klein, and Pryor (2002) define fitting the model as using our data to estimate the unknown parameters $\beta_1, \beta_2, \ldots, \beta_j$. Hosmer and Lemeshow (1989) state that given $(x_i, y_i), i = 1, 2, \ldots, n$, we use the likelihood function to determine the unknown parameters. The likelihood function expresses the probability of the observed data as a function of the unknown parameters (i.e., the vector $\beta_0, \beta_1, \ldots, \beta_k$) and helps us fit the model (Christensen, 1997; Hosmer and Lemeshow, 1989; Kleinbaum, Klein, and Pryor, 2002). The maximum likelihood estimators of these parameters are chosen to be those values which maximize this function. Thus, the resulting estimators are those which agree most closely with the observed data.

Hosmer and Lemeshow (1989) state that when $Y$ is coded 1 or 0 then the expression for $\pi(x)$ above provides the conditional probability that $Y$ is equal to 1 given $x$. This is described as $P(Y = 1|x)$. The quantity $1 - \pi(x)$ gives the conditional probability that $Y$ is equal to zero given $x$, $P(Y = 0|x)$. For pairs such as $(x_i, y_i)$, where $y_i = 1$ the contribution to the likelihood function is $\pi(x_i)$ and for those where $y_i = 0$ the
contribution is \( 1 - \pi(x_i) \), where the quantity \( \pi(x_i) \) denotes the value \( \pi(x) \) computed at \( x_i \).

To express the contribution to the likelihood function for the pair \((x_i, y_i)\) it is convenient, as suggested by Hosmer and Lameshow (1989), to use the following equation

\[
\zeta(x_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}.
\]

In this case because the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given above in the expression:

\[
l(\beta) = \prod_{i=1}^{n} \zeta(x_i).
\]

The principle of maximum likelihood states that we need to use the value which maximizes the expression above; in other words, our estimate of \( \beta \). However, they propose that a mathematically easier way is to work with the \( \log \) of it, which is called the \textit{log likelihood}, and it is defined as

\[
L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln [1 - \pi(x_i)]\}.
\]

In order to find the value of \( \beta \) that maximizes \( L(\beta) \), one differentiates \( L(\beta) \) with respect to \( \beta_0 \) and \( \beta_1 \), the resulting expression to zero as

\[
\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0
\]

and

\[
\sum_{i=1}^{n} X_{ij} [y_i - \pi(x_i)] = 0
\],

where \( j = 1, 2, \ldots, p \), and these equations are called the likelihood equations.
Hosmer and Lameshow (1989) explain that these expressions are nonlinear in $\beta_0$ and $\beta_j$, and thus require special methods for their solution. The methods are iterative in nature. The value of $\beta$ given by solving the likelihood equations is called the maximum likelihood estimate and is denoted $\hat{\beta}$ where the symbol $\hat{\cdot}$ denotes the maximum likelihood estimate of the respective quantity (Hosmer and Lemeshow, 1989; Kleinbaum, Klein, and Pryor, 2002). As a result, $\hat{\pi}(x_i)$ is the maximum likelihood estimate of $\pi(x_i)$. This quantity provides an estimate of the conditional probability that $Y$ is equal to 1, given that $X$ is equal to $X_i$. As such, it represents the fitted or predicted value for the logistic regression model. A consequence of the likelihood equation is that the sum of the observed values of $Y$ is equal to the sum of the predicted values.

Because standard errors of the estimated coefficient are important for the calculation of coefficient testing and confidence interval estimation, it is worth noting here that the method for estimating the variances and covariances of the estimated coefficients are obtained from the matrix of second partial derivates of the log likelihood function (Hosmer & Lemeshow, 1989) as follows:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i)$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} x_{ij} x_{iu} \pi_i (1 - \pi_i)$$

where $j, u = 0, 1, 2, \ldots, p$ and $\pi_i$ denotes $\pi(x_i)$. The inverse of the $(p + 1)$ by $(p + 1)$ information matrix containing the negative of the terms of the equations above (i.e., $\Sigma(\beta) = I^{-1}(\beta)$) provides the variances and covariances of the estimated coefficients.
These terms are denoted as $\sigma^2(\beta_j)$ for the $j$th diagonal element of the matrix (i.e., the variance of $\hat{\beta}_j$, and $\sigma(\beta_j, \beta_u)$ denotes the arbitrary off-diagonal element (i.e., the covariance of $\hat{\beta}_j$ and $\hat{\beta}_u$). Hosmer and Lemeshow (1989) derive the estimators of the variances and covariances (i.e., $\hat{\Sigma}(\hat{\beta})$ by evaluating $\Sigma(\beta)$ at $\hat{\beta}$). The authors use $\hat{\sigma}^2(\beta_j, \beta_u), j, u = 0, 1, 2, \ldots, p$, to denote the values of the matrix and the estimated standard errors as:

$$\hat{SE}(\hat{\beta}_j) = [\hat{\sigma}^2(\hat{\beta}_j)]^{1/2},$$

where $j = 0, 1, 2, \ldots, p$.

Testing for the significance of the coefficients

Hosmer & Lemeshow (1989) state that after estimating the coefficients, what follows is an assessment of the significance of the variables in the model which usually involves formulation and testing of a statistical hypothesis to determine whether the independent variables in the model are significantly related to the dependent variable.

According to Hosmer and Lameshow (1989), an approach to testing the significance of the coefficient of a variable in any model relates to the following question: Does the model that includes the variable in question tell us more about the DV than does a model without it? This question is answered by comparing the observed values of the DV to those predicted by each of the two models (Hosmer & Lemeshow, 1989). The authors explain that the mathematical function used to compare the observed and predicted values depends on the particular problem. If the predicted values with the variable in the model are better or more accurate than when the variable is not in the model, then we determine that the variable in question is significant. When doing this,
they state, we are not considering the question of whether the predicted values are an accurate representation of the observed values in an absolute sense.

The guiding principle with logistic regression, according to Christensen (1997), Hosmer & Lemeshow (1989), Kleinbaum, Klein, and Pryor (2002), and Vittinghoff (2005) is the same as the one for linear regression: compare observed values of the DV to predicted values obtained from models with and without the variable in question. In the logistic regression analysis, the comparison of observed to predicted values is based on the log likelihood function defined above. It is helpful to think of an observed value of the response variable as also being a predicted value resulting from a saturated model, which is a model that contains as many parameters as there are data points.

Hosmer & Lemeshow (1989) provide the following expression to base this comparison

\[ D = -2 \ln \left( \frac{\text{likelihood of current model}}{\text{likelihood of the saturated model}} \right). \]

The quantity inside the brackets in the expression above is called the likelihood ratio and \(-2\ln\) is a mathematical way to obtain a quantity whose distribution is known and can be used for hypothesis testing purposes. The test is called the likelihood ratio test and is represented as follows

\[ D = -2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{\hat{y}_i}{\hat{\pi}_i} \right) + (1 - y_i) \ln \left( \frac{1 - \hat{y}_i}{1 - \hat{\pi}_i} \right) \right], \text{where } \hat{\pi}_i = \pi(X_i). \]

The statistic \( D \) is called the deviance and plays an analogous role in logistic regression as the residual sum of squares plays in linear regression. To assess the significance of an independent variable, one compares the value of \( D \) with and without the IV in the equation. The change in \( D \) due to the absence or presence of the IV in the model is obtained as
and it follows a $\chi^2$ distribution with $p$ degrees of freedom (i.e., the difference in $df$ between the nested model) (Hosmer and Lemeshow, 1989; Vittinghoff, 2005). Vittinghoff (2005) define it as twice the difference between log-likelihoods from the two models and states that it can be referred to the $\chi^2$ distribution for significance testing. The authors explain that because the likelihood for the larger model must be larger than the likelihood for the smaller or nested model, the difference will always be positive.

Hosmer & Lemeshow (1989) state that the result plays the same role the numerator of the partial $F$ test plays in the linear regression. The authors state that because the likelihood of the saturated model is common to both values of $D$ being differentiated to compute $G$, it can be re-written as:

$$G = -2\ln \left[ \frac{\text{likelihood without the variable}}{\text{likelihood with the variable}} \right].$$

In the multivariate case the values are based on the vector $p + 1$ parameters (i.e., the $\hat{\beta}$s). When the $p$ slope coefficients for the covariates in the model are equal to zero, the distribution of $G$ is $\chi^2$ with $p$ degrees of freedom. If the value is equal to or greater than $\chi^2$ the conclusion derived from rejecting the null hypothesis is that at least one or all $p$ coefficients are different from zero.

In order to determine which regression coefficients are statistically significant, Hosmer and Lameshow (1989) and Vittinghoff (2005) recommend looking at the Wald test statistics which are given by the statistical program. This test statistic is denoted as

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}.$$
If the assumption is that an individual coefficient is zero, these statistics will follow the standard normal distribution. The values of these statistics will give an indication of which of the variables in the model may or may not be significant (Hosmer & Lemeshow, 1989, Vittinghoff et al., 2005). For a significance level of approximately .05, the Wald statistic should be above a value of 2. Any variable with a Wald statistic lower than 2 is said to be not significant.

**Multilevel Logistic Regression**

Gelman and Hill (2007) state that multilevel modeling is applied to logistic regression and other generalized linear models in the same way as with linear regression. In other words, they state that the coefficients are grouped into batches and a probability distribution is assigned to each batch. They also explain it as error terms being added to the model corresponding to different sources of variation in the data. A step-by-step description of how data are analyzed for a multilevel logistic regression is given in Chapter 4; however, the motivation for using this type of analysis as outlined by Gelman and Hill (2007) express that the motivations for multilevel modeling are (1) *Learning about treatment effects that vary* which reflects on one of the basic goals of regression analysis which is estimating treatment effects and how the dependent variable changes when the independent variables vary holding everything else constant. The authors state that multilevel models allow us to study the effects that vary by group too. (2) *Using all the data to perform inferences for groups with small sample size*, which refers to the fact that multilevel modeling allows the estimation of group averages and group-level effects. (3) *Prediction* of new cases is a what regression models are commonly used for. (4) Multilevel models allow for the *analysis of data (structured data)* that are collected with an inherent multilevel structure, for example, speech from various speakers. Multilevel
modeling is a direct way to include indicators for clusters or groups at all levels of a design. (5) It is a more efficient inference for regression parameters because data often arrive with multilevel structure, especially in sociolinguistics. (6) These models allow for the inclusion of predictors at two different levels which provides a coherent model that simultaneously incorporates both individual- and group-level models. Finally, (7) these models accurately account for uncertainty in prediction and estimation.

Although these motivations were designed to reflect studies in the social sciences, they fit very well to the needs of the linguist who works with empirical data on language use and they served as the motivations for conducting a multilevel logistic regression as the analysis of my data. First, data collected for a sociolinguistic study is hierarchical by nature because a particular speaker produces many different tokens of the dependent variable, and this speaker belongs to a particular social class, education level, or speech community. Second, the purposes of sociolinguistic studies is to predict the behavior of a specific linguistic structure, in this case copula choice, and its relationship to both the linguistic and group level. Third, using a more updated model (i.e., multilevel model) may help to better predict what social, linguistic, pragmatic and discourse factors help predict copula choice + adjective in Spanish.

**Polychotomization of continuous variables**

The third methodological issue addressed during the design stages of this dissertation was how to address language change through the study of different speakers without jeopardizing the power of the study and yet being able to account for differences due to age.

Labov (1972) put forward three steps for the study of linguistic change in progress. These three steps are the following:
The transition problem is to find the route by which one stage of a linguistic change evolved from an earlier stage. The embedding problem is to find the continuous matrix of social and linguistic behavior in which the linguistic change is carried. The evaluation problem is to find the subjective (or latent) correlates of the objective (or manifest) changes which have been observed (pp. 161-162).

Labov (1972) also states that the value of an explanation rises in relation to its generality, but only to the extent that it rests upon a foundation of reliable and reproducible evidence (p. 162). To search for this type of evidence, he states that the simplest data that can establish the existence of a linguistic change is a set of observations of two successive generations of speakers. These generations should be of comparable social characteristics that represent stages in the evolution of the same speech community (p. 163). Finally, Labov (1972) observes that solutions to transition problems (due to the observation of two generations) depend upon close analysis of the distribution of linguistic forms along the dimensions formed by the age groups of the population (i.e., apparent time). This concept has been operationalized as means of analysis in the field of sociolinguistics by grouping speakers into age groups (Aguilar-Sánchez, 2007; Bentivoglio & Sedano, 1993; Díaz-Campos & Geeslin, in press; Fishman, 1972; Gutiérrez, 1994b; Labov, 1972; Silva-Corvalán, 1994) creating what statisticians called polychotomization of a continuous variable (Maxwell & Delaney, 1993).

Polychotomization is detrimental to the power of regression analysis (Aguinis, 1995) and it can lead to false statistical significance and the overestimation of strength of relationships accompanied by an increase in Type I errors (Maxwell & Delaney, 1993). Maxwell and Delanay (1993) advocate for avoidance of this practice because it affects the power of the study and it leads to possible spurious conclusions. The authors state that
there is a quite high potential for misinterpretation because many of the measures studied this way are without a doubt highly correlated.

Even though Labov (1972) advocates for the study of language change through the study of three different generations and that the practice is to polychotomize the variable age to reflect this suggestion, I followed Maxwell and Delaney’s (1993) suggestions and avoided polychotomizing and proceeded to set the level of power and to calculate appropriate sample size for this study. All these variables are analyzed to capture nuances originating from the diglossic situation present in Limón as well as the amount of contact present in this context.

Following this assumption, the present dissertation will be based on the concept of apparent time and will be conducted by taking into account the different generations of bilingual speakers in the sampling of Limonese English and the varieties of Spanish found in Costa Rica; however, age will be regarded as a continuous rather than a grouping variable. Because age is taken as a continuous variable, any difference due to age is provided if the variable “age” turns out to be influential in the prediction of copula choice in Spanish. In other words, no grouping regarding age is done for the data in the present study.

**Sample Size**

In order to lay out the rationale for the selection of the participants for this research, a discussion on issues related to the calculations of sample size is presented to justify the number of subjects and their distribution selected for this dissertation. The section is structured with a discussion of the importance of proper sample size calculation for achieving a specific power, a discussion of sample size calculation for the specific statistical analysis used in this dissertation, and the presentation of the results of two a
priori Monte Carlo simulations conducted to determine the sample size needed to achieve a power level of .8. A power of .8 means that there is only 80% chance of rejecting the null hypothesis when it should be rejected, in other words finding significant results when they should be found.

**Background**

In recent years, sample size planning has become a very important issue among researchers of the social sciences. It is important because sample size planning is one of the key factors to consider when designing a research study (Kelley & Maxwell, 2008) and it is related to the power of the study. In psychology, reporting the calculation of sample sizes as part of research design has become part of the suggestions for the publication of studies which states that, as suggested by Wilkinson and the APA Task Force on Statistical Inference (1999), each study should provide information on sample size and the process from which it was derived. Because of its relationship to the power of a study, an appropriate sample size becomes one of the most valuable elements in the design of any research.

Traditionally, appropriate sample size calculations in the social sciences have been approached through a power analytic framework. The power of a study is closely related to its replicability, which results in the building of a body of cumulative knowledge. In order to achieve this accumulation of knowledge, studies should have a high level of power, for which sample size is a key element (Cohen, 1988). Wilkinson and the APA Task Force on Statistical Inference (1999) also suggests that power computations are most meaningful when they are done before data are collected and examined as part of the study design. It is also important to show how effect-size estimates, a key element on the calculation of sample sizes, have been derived from
previous research and/or theory in order to dispel suspicions that they might have been constructed to justify a particular sample size (p. 596).

Despite the importance given to the statistical significance of a test, very little attention has been paid to the report of the calculation of sample sizes in the field of linguistics, let alone to the power of each study. Because proper sample size calculation leads to a stronger study and to more generalizable results, I have decided to dedicate part of this chapter to the discussion of the importance of the appropriate of sample size and how it was calculated for this dissertation. It is the goal of this dissertation to follow the standards that have helped other fields accumulate a cohesive body of knowledge and improve practices in linguistic studies.

Importance of proper sample size calculations

Kraemer & Thiemann (1987) state that the scientific method requires researchers to propose a theory and put it to empirical tests. They also postulate that theory is considered false until demonstrated beyond reasonable doubt to be true, and that this is expressed as an assumption that the null hypothesis is true. Thus, a researcher evaluates the preliminary evidence and decides whether or not the case is important enough to test the hypothesis. The preliminary evidence consists of the literature review, case histories, theoretical considerations, and pilot studies, among others. Then, for a test to be a valid 5% or 1% test, the computation must be correct 95% or 99% of the times. In other words, a researcher needs to be sure that the analysis s/he is conducting has a desired probability of being correct. The procedures used to analyze data should be designed to appropriately account for the description of the data; therefore, the right test for the right set of data must be a priority. Furthermore, Leon (2004) states that the study must be designed with a sample size that is adequate to detect the differences, in this case levels of association,
that are hypothesized. Good (2005) compares sample size with a digital photograph when he states that the ability to see greater detail is a “squared function” because it takes four times as large a sample to see twice the detail. He also states that one can get the general picture “cheap”, but precision is “costly”. However, this should not be confused or generalized to the common notion of any large number of subjects or observations is as good as a well planned sample size for the specific study being designed.

Cohen (1992b) explains that in planning research, deciding the sample size is crucial because research costs are approximately linear in the number of subjects. Therefore cost-effectiveness demands that this decision be appropriate. Sprinthall (2003) postulates that one issue that researchers face is how large the group must be to provide a certain level of confidence that the selected sample comes as close as possible to truly mirroring the population while Lenth (2001) defends the idea that the study must be of adequate size relative to the goals of the overall study. Lenth (2001) states that the sample must be ‘big enough’ that an effect of such magnitude to be of scientific significance can also be statistically significant. Even though both authors voice two different ideas of the importance of the size of the sample, both ideas lead to the same conclusion: calculation of sample size prior to the collection of data is an important step in any research design.

In any research field, it is often advocated that, to improve the power of a study, one must increase the sample size (Cohen, 1992b). Lenth (2001) wisely calls for caution in this regard and he adds that it is also important that the study not be ‘too big’ where an effect of little scientific importance is easily statistically detectable. He also states that sample size is important for economic reasons, a point that agrees with Cohen’s (1992a) claim on the importance of proper power calculations. Cohen (1992a) postulates that an
undersized study can be a waste of resources for not having the capability to produce useful results, and an oversized one uses more resources than are necessary.

Lipsey (1990) asserts that sample size calculation is important because virtually any desired level of power can be attained in any design simply by making the sample size large enough. However, for the range in which most treatment effect sizes fall or, in this case, the levels of association, the sample size needed to attain high power levels is often much larger than what is customary under most research circumstances. Furthermore, Lipsey (1990) brings the discussion beyond the question of how many subjects are needed. He states that other than the question of how many subjects, we need to ask about their distribution across experimental conditions in a research design. We need to take into account that the relationship between statistical power and sample size is based less on the total number of subjects involved than on the number in each group or cell within the design. Thus, close attention must be paid to the effect of the number of groups over which subjects are distributed and the proportion of subjects within each group. In the case of this dissertation, this statement can be translated as the effect of the number of tokens (i.e., observations) per speaker and the number of speakers (i.e., groups of observations) needed to achieve the desired power as presented at the end of this chapter.

Approaches to sample size determination

After discussing the importance of carefully planning the number of subjects to employ in a research design, it is evident that most authors approach sample size as a component to determine the power of a test. This tendency derives from Cohen’s (1988) definition of sample size as the invariant feature of sample precision and as one of the elements used to determine the power of a test.
More recently, sample size planning has been approached as a way to accurately estimate the population parameters (Kelley, 2008; Kelley & Maxwell, 2003, 2008; Kelley, et al., 2003; Kelley & Rausch, 2006). Even though for this dissertation I have decided to follow a sample size planning for power, both current approaches are described below because they are part of the decision making process.

Kelley & Maxwell (2008) assert that an appropriate sample size depends on the goals of the researcher. These goals could be related to one of the two approaches mentioned above or to both. They state that there are several possible scenarios depending on the goal of the researchers. Figure 3.2 (taken from Kelley & Maxwell, 2008) provides a summary of these scenarios.

<table>
<thead>
<tr>
<th>Goal</th>
<th>Accuracy</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.2. Two by two conceptualization of possible scenarios when statistical power is crossed with statistical accuracy.

In this context the researcher can be interested in the overall effect size of the best-fitted model (i.e., context a), or s/he can be interested in the accuracy of the parameter estimation from the sample (i.e., context c). S/he can also be interested in both power and accuracy (i.e., contexts a & c). If the researcher is interested in the particular effect that one of the variables has in the regression model, s/he can select the power of this particular variable (i.e., context b). S/he can also be interested in the accuracy of the estimation of the effect of this variable (i.e., context d), or in both accuracy and power of that particular variable (i.e., contexts b & d). Similarly, Zodpey (2004) proposes that two
of these approaches can be used to draw statistical inference from study results. One is hypothesis testing (i.e., power) and the other is accuracy estimation (i.e., confidence interval approach). Kelley and Maxwell (2008) point out that, given the goals of the researcher, any combination of the cells in the table is possible and that multiple cells in Figure 3.2 can be relevant. In the present work I am concerned with contexts \( a \) and \( b \) because the goal of this section is the determination of the appropriate sample size to achieve a power of .8 for both the omnibus effect (i.e., the saturated model) and for a specific parameter estimate (i.e., “gradiency”).

Following the suggestions laid out by Wilkinson and the APA Task Force on Statistical Analysis (1999), I chose sample size planning for power during the design stage of this study and the presentation of confidence intervals for the report of the results after data are analyzed.

**Determination of Sample size for the study of copula + adjective in Limon, Costa Rica**

In order to fulfill the requirements for research design postulated above, and to achieve a power of .8 (i.e., to have an 80% probability of rejecting the null hypothesis when the null hypothesis is not true in the population), the following procedure was followed to achieve the appropriate sample size for this dissertation.

**The statistic test**

As explained above, the purpose of this dissertation is to predict how copulas are chosen in the context of \( \text{copula} + \text{adjective} \). My data structure is hierarchical, and a multilevel logistic regression is the most appropriate statistical analysis to be employed despite the tradition in sociolinguistics for using multivariate logistic regression. Multivariate logistic regression assumes that the data comes in a non-hierarchical/not grouped manner; therefore it is not appropriate for data collected for a sociolinguistic
study. Sociolinguistic data is hierarchical in nature (i.e., speakers produce X number of tokens, speakers belong to X group, etc.); therefore, it should be analyzed as such.

In order to fulfill the sample size calculation requirements, a procedure for multilevel logistic regression presented by Gelman and Hill (2007) is followed. In this procedure, data are analyzed using the statistical software R (R Development Core Team, 2009) and the package “arm” (Gelman, Su, Yajima, Hill, Gracia Pittau, Kerman & Zheng, 2009). The statistical software R is an open-source software and is widely used for advanced statistical analysis due to its flexibility for modeling. From the package “arm”, I use the function “lmer()” stands for “linear mixed effects in R”. This function works for generalized linear models (Gelman and Hill, 2007) and yields results at two levels. The first level is the inference about the intercept and slope for the token (i.e., linguistic) variables. The second level of the output displays the estimated variation for the social variables. It reports the deviance statistic at the end. A full and detailed account of how this analysis works will be given in Chapter 4.

The parameters

In order to achieve a power of .8, the alpha level is set to 0.05 following conventions in the field of linguistics and to avoid the loss of power due to stringent alpha levels.

Because effect sizes are hard to calculate for logistic regression and because comparability among previous studies of copula + adjective is not possible due to the differences in design, I have decided to run a priori Monte Carlo simulations to calculate the coefficients and their standard errors for a more accurate calculation of effect sizes. An a priori Monte Carlo simulation is a procedure that simulates the data given structure and performs the analyses thousands of times in order to get an average number of times
a particular coefficient or a model is statistically significant. In this case, the coefficient is said to be significant if it is two standard errors above or below the mean. In other words, alpha is set to .05 and a 95% confidence that the null hypothesis will be rejected.

Furthermore, since Aguilar-Sanchez (2007) found that on average speakers produce 30 instances of ser and estar, the Monte Carlo simulations were set to calculate sample sizes at the observation level of 25, 30 and 35. These numbers were labeled $K$ following Gelman and Hill (2007) (see Appendix 2 for the full code). These numbers correspond to the lowest number of observations found, the average, and the highest number found in the data used in Aguilar-Sanchez (2007). With regards to the number of speakers, the numbers were set to 20, 30, 60, 100, 200, and 300. These numbers were set to come up with a graph that allowed me to determine the appropriate sample size needed to design a study with a power level of .8. They were labeled $J$ for simulation purposes.

**The a priori Monte Carlo Simulations**

The first a priori Monte Carlo simulation was set to look for the sample size appropriate to find different levels of power for the omnibus effect of the saturated model (see Appendix 1). In other words, the effect of all the variables included in the model. Table 3.2 shows the results of these calculations.
Table 3.2

Power Calculations for the Omnibus Effect of the Saturated Model

<table>
<thead>
<tr>
<th>number of observations</th>
<th>number of speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K</td>
</tr>
<tr>
<td>25</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

As Figure 3.3 shows, a sample size of 30 observations and around 60 speakers is needed to achieve a power of .8. As we see from this table as the number of speakers ($J$) increases, the power of the study increases. In a like manner, increasing the number of tokens per speaker will produce an increase in power.
The second a priori Monte Carlo simulation was conducted with the same parameters as the first one. The only difference between the two was that for the second I decided to look for the sample size appropriate to detect significant effects regarding the variable *gradiency*. This variable turned out to be of interest in the study of *copula + adjective* in Costa Rican Spanish (Aguilar-Sánchez, 2007).

Table 3.3 and Figure 3.4 show the results of the Monte Carlo simulation (see Appendix 2 for partial R code).

Table 3.3 Power Calculations for the Targeted Effect of the Variable Gradiency

<table>
<thead>
<tr>
<th>number of observations</th>
<th>number of speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>J</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td>K=25</td>
<td>0.58</td>
</tr>
<tr>
<td>K=30</td>
<td>0.7</td>
</tr>
<tr>
<td>K=35</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Figure 3.4. Sample Size Calculations for the Targeted Effect "Gradiency".
As we can see, similar to the previous analysis, as sample size is increased, the power of the study increases and the more tokens are included, the higher the level of power. As for the general model, approximately 60 speakers with an average of 30 observations are needed to achieve a power of .8 for this study when one is interested in finding significant results for the variable *gradiency*.

The decision of determining how many interviews to collect was derived taking into account three aspects. First, the desired power for the full model and for the targeted effect “*gradiency*”. Second, because in the pilot study speakers produce an average of 30 instances of copula, it is said that 30 tokens is the population parameter for this calculation. 35 observations brings power calculations to the desired level sooner than 30 in both simulations; however, the population parameter is what helps make the decision to select the power yielded by the average number of observations found in the population and to avoid problems of losing power for small response probabilities. And third, even though more speakers will bring a higher power to the study, a level .8 is the standard for the behavioral sciences and it has been demonstrated to be the appropriate level for both scientific and logistic reasons.
Chapter 4

Methodology

Introduction

Chapter 4 presents a description of the methodological design of this investigation. This chapter is divided into five sections that provide the details and rationale followed for the decision-making regarding each component of the research design. The first section restates the research questions that guided the present study. The second section includes the data collection instruments and the rationale for selecting them, and data collection procedures. The third section introduces the data coding scheme as well as the data coding procedures for the present work and how they fit into the bigger picture of research on copula + adjective in Spanish. The fourth section of this chapter contains a description of the sample size selected for this dissertation based on the results of the Monte Carlo simulations described in Chapter 3. The last section presents the rationale for and description of the statistical test used for the analysis of the data collected for this study.

The Research Questions

After reviewing the relevant literature and the theoretical accounts regarding copula + adjective in Spanish and after addressing key methodological issues regarding power of the study and sample size determination, I proceed to answer the following questions:

1. What linguistic variables predict the use of ser and estar + adjective in the Spanish spoken by Costa Ricans in Limón?
2. What social variables predict the use of ser and estar + adjective in the Spanish spoken by Costa Ricans in Limón?
3. Can the pattern of variation in the use of ser and estar be considered a change in progress or a stable change?
The design of this dissertation and the rationale behind it are presented in the following sections.

**The Sample**

To achieve a power of .8 (i.e., to have an 80% probability of rejecting the null hypothesis when the null hypothesis is not true in the population) with a p-value set at 0.05 for all analyses, the sample size for this study needed to be around 60 speakers. A total of 58 speakers participated in this corpus (See Table 4.1). Their ages ranged between 16 and 91 years old. Twenty-six were males and thirty-two females. All have completed at least three years of primary education. Eighteen were monolingual speakers of Limonese Spanish, thirty-five spoke two languages, and five reported speaking at least three languages. Nineteen have restricted contact with English and thirty-nine reported non-restricted contact with English. This categorization is explained below. All but three reported a positive attitude towards keeping indigenous languages such Limonese English. Five were retired, one unemployed, thirty one work in services, six at home, eight in education, one in land and six were students.
Table 4.1

Distribution of Speakers by Social Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>26</td>
<td>44.83</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>32</td>
<td>55.17</td>
</tr>
<tr>
<td>Education</td>
<td>Primary incomplete</td>
<td>2</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>Primary complete</td>
<td>5</td>
<td>8.62</td>
</tr>
<tr>
<td></td>
<td>Secondary incomplete</td>
<td>11</td>
<td>18.97</td>
</tr>
<tr>
<td></td>
<td>Secondary complete</td>
<td>15</td>
<td>25.86</td>
</tr>
<tr>
<td></td>
<td>University incomplete</td>
<td>7</td>
<td>12.07</td>
</tr>
<tr>
<td></td>
<td>University complete</td>
<td>18</td>
<td>31.03</td>
</tr>
<tr>
<td>Bilingualism</td>
<td>Monolingual</td>
<td>35</td>
<td>60.34</td>
</tr>
<tr>
<td></td>
<td>Bilingual</td>
<td>5</td>
<td>8.62</td>
</tr>
<tr>
<td></td>
<td>Multilingual</td>
<td>18</td>
<td>31.03</td>
</tr>
<tr>
<td>Contact</td>
<td>Restricted</td>
<td>19</td>
<td>32.76</td>
</tr>
<tr>
<td></td>
<td>Not-restricted</td>
<td>39</td>
<td>67.24</td>
</tr>
<tr>
<td>Attitude</td>
<td>Important to save languages</td>
<td>55</td>
<td>94.83</td>
</tr>
<tr>
<td></td>
<td>Not important to save languages</td>
<td>3</td>
<td>5.17</td>
</tr>
<tr>
<td>Profession</td>
<td>Unemployed</td>
<td>1</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>Services</td>
<td>31</td>
<td>53.45</td>
</tr>
<tr>
<td></td>
<td>Home</td>
<td>6</td>
<td>10.34</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>8</td>
<td>13.79</td>
</tr>
<tr>
<td></td>
<td>Land</td>
<td>1</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>Retired</td>
<td>5</td>
<td>8.62</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>6</td>
<td>10.34</td>
</tr>
</tbody>
</table>
Data collection

Based on Ladegaard’s (2000) call for the use of eclectic methods that include qualitative and quantitative research and on Hyrkstedt & Kalaja’s (1998) call for a change in the collection of data and its analysis; I have decided to use a data collection method that includes both a background questionnaire with qualitative and quantitative data and discourse data via Labovian interviews. This was done in order to collect a corpus that allows for generalization of findings. Each instrument is described in full detail in the following sections. The procedures for the collection of data for each instrument follow their descriptions.

Instruments

Informed Consent

The first document given to subjects was the informed-consent form approved by the Human Subjects Committee at Indiana University. This form explained the purposes of the research without revealing the linguistic structure under study. The study was registered under protocol #07-12218 of the Indiana University Human Subjects Committee Review.

Background Questionnaire

The background questionnaire is a series of 12 open-ended questions that elicited background information, as well as 1 yes/no question that elicited attitudes towards each of the languages spoken in Limón and a space for justifying the answer. The description of each of these questions is presented below. Each version of the background questionnaire presented hereafter is an English translation of the original written in Spanish.
Figure 4.1 presents the background questionnaire, which was designed in three sections. The first one is related to demographic information, the second is related to language use and literacy, and the third section is related to language attitudes. Each of these questions is presented below.

Questions one through eight elicited information on each participant's age, gender, economic status, place of residence and for how long they have lived there, and languages spoken. Economic status was gathered by asking questions regarding education and profession due to the fact that in Costa Rica social and economic status can be determined by those factors since professional opportunities and salaries are tied to educational level. The minimum wage per profession and per degree obtained is government regulated (Inman, Mesa, Oleas, & de los Santos, 1997).

Questions nine through twelve were designed to elicit the amount of contact with English or Spanish, via open-ended questions. The latter is the official language of the country, and the former is the language under inquiry in this question. Operationalization of the level of contact will be discussed in the section related to the coding of data in a subsequent section.
The last question presented to the participants was related to choices and attitudes towards the indigenous languages spoken in Limón. This variable has proven to be significant in the analysis of language in this region and with regard to other languages because it elicits identity issues related to the preservation or decline of the use of a

---

**Background Questionnaire**

Instructions: Please answer the following questions with your information, and give your opinion where appropriate.

**Code:** ___________________________________________

1. **Gender:** M  F  2. **Age:** __________

2. **Education:** ___________________________________________ (last level achieved)

3. **If Student, what do you study?** ___________________________________________
   **Mother’s profession:** ___________________________________________
   **Father’s profession:** ___________________________________________

4. **If not a student: Profession:** ___________________________________________

5. **Place of birth:** ___________________________________________
   **How long did you live there?** ___________ years.

6. **Where do you live?** ___________________________________________ **For how long?** ___________

7. **What languages do you speak?** Spanish  English  Chinese  Other
   **If other, which one?** ___________________________________________

8. **At what age did you learn to speak Spanish?** ___________ English? ___________

9. **Do you write Spanish?** Yes  No  Somewhat  English? Yes  No  Somewhat

10. **Do you express yourself best in Spanish, English, or both?** ___________________________________________

11. **With whom do you speak Spanish the most?** English? ___________________________________________

12. **Where do you speak Spanish the most?** English? ___________________________________________

13. **Do you think that it is important to preserve the English spoken in Limón as well as the indigenous languages?** Yes  No
   **Why?** ___________________________________________

---

**Figure 4.1. Background Questionnaire**

The last question presented to the participants was related to choices and attitudes towards the indigenous languages spoken in Limón. This variable has proven to be significant in the analysis of language in this region and with regard to other languages because it elicits identity issues related to the preservation or decline of the use of a

**Labovian Interviews**

The tokens for this project were collected using the typical Labovian sociolinguistic interview (Labov 1972, Tagliamonte 2006). These interviews were held over a period of nine months starting in November 2007 and ending in July 2008.

<table>
<thead>
<tr>
<th>Interview Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Greeting</td>
</tr>
<tr>
<td>2. Explanation of what I am doing</td>
</tr>
<tr>
<td>3. Explanation of the informed consent form and of issues related to confidentiality</td>
</tr>
<tr>
<td>4. Explanation of how the recording works.</td>
</tr>
<tr>
<td><strong>5. Interview</strong></td>
</tr>
<tr>
<td>5.1. Could you describe your childhood in full detail?</td>
</tr>
<tr>
<td>5.2. How does the city you grew up in compare to the city you live in right now?</td>
</tr>
<tr>
<td>5.3. Could you describe in full detail how you see life in this city?</td>
</tr>
<tr>
<td>5.4. Could you tell me what things you would change and why?</td>
</tr>
<tr>
<td>5.5. Tell me what you like the most about the Caribbean culture and why?</td>
</tr>
<tr>
<td>5.6. Can you describe any of the activities that characterize the Caribbean culture you just described?</td>
</tr>
<tr>
<td>5.7. Can you tell me how you see the language situation in Limon?</td>
</tr>
<tr>
<td>5.8. Could you describe in full detail how you see Limon in 20 years?</td>
</tr>
<tr>
<td>5.9. Can you tell me what you think causes social problems in Limon and why?</td>
</tr>
<tr>
<td>6. End of recording</td>
</tr>
<tr>
<td>7. Thank the participant and give the compensation</td>
</tr>
</tbody>
</table>

Figure 4.2 shows the interview questions used. They were based on questions regarding cultural traditions including food, dance, music, language use, childhood memories, school years stories, current social and political situation, and significant personal experiences. These interviews have been used in the field of sociolinguistics and
they have proven to be excellent sources of linguistic data (Aguilar-Sánchez, 2007; Bentivoglio & Sedano, 1993; Díaz-Campos, 2003; Díaz-Campos & Geeslin, in press; Labov, 1972; Labov & Dingwall, 1978; Sankoff, 1988a; Silva-Corvalán, 2001).

**Procedure**

Each interview was conducted following standard interview procedures to guarantee obtaining speech samples that reflect the language commonly used in the community in daily situations. With the aid of a community contact (i.e., a second interviewer), each participant was approached by the researcher and asked whether they could spare thirty minutes of their time. Instruments were provided to each subject who agreed to participate, starting with the informed consent form. In this form, participants were told they were participating in a study to collect historical as well as current information on the province of Limón. Subjects were told that their participation was voluntary and that they could withdraw from the study at any time. No subjects withdrew from the study. They were also assured that their identity would be strictly confidential when reporting the results.

Each participant initialed, dated and signed the informed consent form and kept a copy for their records. For participants who could not read or write, the researcher was in charge of reading the form aloud for them to sign and initial.

After signing and initialing the consent form, participants were presented with the background questionnaire. In order to keep consistency of the information gathered through this instrument, the researcher asked the questions and filled out the questionnaire himself. This part of the interview was not recorded.

Once the background questionnaire was filled out, the researcher explained how the voice recorder works and advised the participant to speak clearly. The researcher then
positioned the recorder in a way to capture the participant’s voice. The interviews were conducted in a variety of locations from offices and classrooms to public places such as restaurants, parks, or the beach.

Each interview was conducted following sociolinguistic interview procedures and was recorded using an Ipod Photo 60GB and an Ipod microphone or a Sony ICRecorder MP3 (ICD-U70). The researcher started with a question and asked follow up questions to keep the informant talking. The researcher tried to minimize his speaking time by acknowledging understanding of the message and by asking short follow up questions. Each major question asked (See Figure 4.4 above) was followed by normal conversation assent signs such as “umhum”, “de verdad?” seriously, “que bien” good, among others. If the participant was not very talkative the researcher asked more questions to elicit enough data. At the end of each interview the researcher made sure that the interview had been recorded, then thanked the participant and provided him/her with a small compensation.

Data coding

Data coding scheme

Data coding for the present investigation was done following Diaz-Campos and Geeslin’s (in press) and Aguilar-Sanchez’ (2007) coding schema, with slight modifications, for instances of copula + adjective. The term copula is used to represent instances of ser and estar only. All other copular verbs (e.g., parecer ‘seem’) are excluded. Aguilar-Sanchez’ (2005a) scales of socioeconomic status and contact level were used to assure comparability between studies and for generalization of the findings to the field of study. Variables from more recent studies on copula choice were also included and are introduced below.
Table 4.2
Coding of Independent Linguistic Variables at the Sentence Level for the Studies of Copula Choice: The Sentence

<table>
<thead>
<tr>
<th>Variable</th>
<th>Categories</th>
<th>Example</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Sentence</td>
<td>[+ stage-level]</td>
<td>Hoy Elena está enferma [Today Elena is sick]</td>
<td>Is the interpretation limited in time?</td>
</tr>
<tr>
<td></td>
<td>[- stage-level]</td>
<td>Elena es amable [Elena is polite]</td>
<td></td>
</tr>
<tr>
<td>Susceptibility to change</td>
<td>[+ susceptible]</td>
<td>El niño es bajo [The boy is short]</td>
<td>Can the quality of the referent change?</td>
</tr>
<tr>
<td></td>
<td>[- susceptible]</td>
<td>La montaña es alta [The mountain is tall]</td>
<td></td>
</tr>
<tr>
<td>Experience with the referent</td>
<td>[indirect]</td>
<td>En Limón los carnavales son bonitos [In Limón the festivals are nice]</td>
<td>Does the speaker have first-hand knowledge?</td>
</tr>
<tr>
<td></td>
<td>[ongoing]</td>
<td>El problema de la basura es preocupante [The garbage problem is worrisome]</td>
<td>Is it on-going perception?</td>
</tr>
<tr>
<td></td>
<td>[immediate]</td>
<td>La sopa está mortal! [The soup is wonderful]</td>
<td>Is it an immediate reaction?</td>
</tr>
<tr>
<td>Adverbs</td>
<td>[+ overt]</td>
<td>Hoy el clima está muy bonito [Today the weather is very nice]</td>
<td>Is the predicate modified by an overt temporal adverb?</td>
</tr>
<tr>
<td></td>
<td>[- overt]</td>
<td>El restaurante está lleno (hoy)! [The restaurant is full (today)!]</td>
<td>Is the predicate modified by a covert temporal adverb recovered from the discourse?</td>
</tr>
<tr>
<td></td>
<td>[absent]</td>
<td>El clima está muy bonito [The weather is very nice]</td>
<td>There is no overt or covert adverb related to the predicate</td>
</tr>
<tr>
<td>Subject</td>
<td>[+ first order]</td>
<td>Juan es joven [Juan is young]</td>
<td>Does the subject referent exist as an entity?</td>
</tr>
<tr>
<td></td>
<td>[- first order]</td>
<td>La situación está mala [The situation is bad]</td>
<td>Is the subject an abstract concept or an idea?</td>
</tr>
</tbody>
</table>
Table 4.2 shows the variables included in this study. The dependent variable is the use of *ser* or *estar*. Independent variables can be divided into three groups. The first group includes variables related to the sentence. The sentence includes the verb and its complements including adjuncts. The decision to include all the parts of the sentence in the analysis comes from Verkuyl (1993, 2001). He advocates the use of all components in a sentence to investigate aspectual composition. It was my decision to follow this practice in order to construct a statistical model using all possible factors that might influence copula choice in Spanish. This decision to include more linguistic features is consistent with the latest methodologies used in the study of copula choice in Spanish. (Díaz-Campos & Geeslin, 2005; Geeslin, 2003; Geeslin & Díaz-Campos, 2005a; Geeslin & Guijarro-Fuentes, 2007; Guijarro-Fuentes & Geeslin, 2006; among others)

The variable *predicate reading* in the present work combines predicate type which examines if a predicate is limited in time as well as frame of reference which examines if a referent is compared to itself at two different points in time or with a group of like referents (Clements, 2006). Example 4.a shows a referent compare with a group of like referents, while 4.b. is a comparison of the same referent at two points in time.

4.a “mi mamá ERA muy abierta y me daba la oportunidad de realizar esas cosas.”

M29D04IVA

‘my mom was very open-minded and gave me the opportunity to do many things’.

It can also be a stage-level reading which compares the referent with him/her/itself at another point in time as in 4.b.

4.b. “antes ESTABAMOS conformes con tener una bicicleta que [risas](los aros de la bicicleta) que no ESTABAN errumbrados.” M37D10IVA

‘before we were happy with having a bicycle even thoug [laughs] (the rims of my bicycle) which were not rusted’.
Individual-level readings are predicted to appear more with *ser* and stage-level reading with *estar*. This new variable comprises two variables that had been used before (i.e., *predicate type* and *frame of reference*), but that as part of the construct might be closely related to one another. My decision to combine these two variables was based on Yoon’s (2001) suggestion as a theoretical foundation and because by doing so I assure that both notions are represented. Of these two separate variables only *predicate type* has been proven to be a strong predictor of copula choice in Spanish. *Frame of reference* has only been statistically significant in one study: Guijarro-Fuentes and Geeslin (2006). With this evidence and because of the theoretical foundations of each variable, I decided to include this modification to guarantee that the statistical model will not have problems with two highly correlated variables.

The next variable is *susceptibility to change* which refers to whether the quality of the referent could change over time as in 4.c

4.c. “porque las cosas ESTÁN tan feas que uno no...” M78A11GIN
   ‘because things are so ugly that one does not…’

or could not change as in 4.d

4.d. “esa casa ES bien vieja, no recuerdo” F91C78IVA
   ‘that house is so old, I don’t remember’.

This variable derives from the notion of change (Bull, 1942; De Mello, 1979) and its operationalization in previous studies. It is included in this study because it has been proven to be a strong predictor of copula choice in almost all studies that included it (see Table 2.4 in Chapter 2).
The next variable is *experience with the referent* which answers the question of whether the speaker has first-hand knowledge of the event. This variable codifies whether this knowledge was ongoing as in 4.e

4.e. “antes (las sopes de mondongo) ERAN bien espesas, buenas” M52C01IVA
‘long ago, tripe soup were very thick, good’

immediate as in 4.f

4.f. “ahora ya (Limon) ESTA muy poblado.” F59A02RCH
‘right now (Limon) is already too populated’

or indirect as in 4.g

4.g. “Vea la cultura de limón mi hija dice que ES, ES muy rica.” F41D08IVA
‘look, my daughter says that Limon’s culture is, is very rich’.

*Estar* is said to appear more when the speaker has an immediate experience with the referent and *ser* with those that are ongoing experiences. Different sociolinguistic studies have operationalized this variable (Díaz-Campos & Geeslin, in press; Geeslin & Guijarro-Fuentes, 2008; Guijarro-Fuentes & Geeslin, 2006) and in most of them it has proven to be a strong predictor of copula choice. It has also helped explain the extension of *estar* in others (Guijarro-Fuentes & Geeslin, 2006). Thus, my decision to include it in the present study.

The next variable is *adverb*. There are three categories within this variable. The first one is the absence of temporal adverb as in 4.h

4.h. “si ustedes piensan que con ese título la van a pasar conmigo, ESTÁN equivocados”. M21000IVA
‘if you think that with that title you are going to pass the course, you are wrong’

the presence of a temporal adverb or temporal expression as in 4.i

4.i. “o SEA, el estado de la calles ahorita ES peor.” F28D11IVA
‘I mean, the condition of streets is worse now’
or the presence of an implied adjective or temporal expression as in 4.j

4.j. como le digo, (el barrio) ERA muy calmado, (antes).” M28C06IVA
‘As I am telling you, (this town) was very calmed (before).’

The presence or absence of an adverb was included because it has been proposed that this feature helps us to classify these two copulas. These proposals come from two different approaches: the generative grammar approach (Schmitt & Miller, 2007) and the functional approach (Clements, 2006; 1988). It was suggested (Ortíz-López, 2000) and it has also been proven to help explain the extension of estar (de Jonge, 1993; Salazar, 2007). The predictions provided by the authors are that the presence of a temporal expression or an adverb pairs with the copular verb estar. In the present work, all types of temporal expressions were coded.

Another independent variable included is subject. This variable distinguishes between first-order-entities (i.e., people, animals, plants, and things) as in 4.k

4.k. “ella, como ERA muy conocida, (antes)”. M68C10IVA
‘because she was well known, (before)’

and second-order-entities (i.e., states, processes, and events, more likely to be located in time and to take place rather than exist) as in 4.1

4.1. “aunque (comer tortuga) ESTÁ prohibido,(ahora).” M32B03IVA
‘even though (eating turtle) is prohibited, (now)’.

Although whether the referent is animate has been significant in the prediction of copula choice only in two studies Silva-Corvalán (1986) and Geeslin and Guijarro-Fuentes (2008); I have decided to include the term subject in this study to refer to the referent based on two issues. The first one is based on the principle of replicability and sampling error. In other words, I search for further evidence that the fact that the subject
is not influential in the selection of copula in Spanish is true and not due to sampling error and the newness of the analysis of this phenomenon in the Spanish of Limon and Costa Rica. In this variety of Spanish, this type of analysis has not been conducted and variables cannot be thrown out due to their behavior in other varieties of Spanish. The second one is theoretical because the operationalization of this variable differs from the others. In previous studies, animacy of the referent has been taken as a variable (Geeslin & Guijarro-Fuentes, 2008; Silva-Corvalán, 1986) and not as a direct representation of the grammatical subject. For the present study I follow Clement’s (2006) classification of subjects into first-order and second-order entities. This classification includes animate and inanimate referents, but it is not equivalent that done in previous studies in that I classify the grammatical subject of the sentence and not just the referent.
Table 4.3

Coding of Independent Linguistic Variables at the Sentence Level for the Study of Copula Choice: Adjectives

<table>
<thead>
<tr>
<th>Variable</th>
<th>Categories</th>
<th>Example</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Adjective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resultant State</td>
<td>[+ resultant]</td>
<td>El restaurante está lleno hoy</td>
<td>Is the adjective a resultant state?</td>
</tr>
<tr>
<td></td>
<td>[- resultant]</td>
<td>El rice’n’beans es barato</td>
<td></td>
</tr>
<tr>
<td>Adjective Class</td>
<td>[age]</td>
<td>Joven [young]</td>
<td>Which semantic class best describes the adjective (in the sense it is used in the given predicate)?</td>
</tr>
<tr>
<td></td>
<td>[size]</td>
<td>Grande [large]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[nationality/orIGIN/religion]</td>
<td>Limonense [Limonese]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[description/evaluation]</td>
<td>Dificil [difficult]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[description of a person(nality)]</td>
<td>Inteligente [intelligent]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[color]</td>
<td>Azul [blue]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[mental/physical state]</td>
<td>Animado [animated]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[sensory characteristic]</td>
<td>Sabroso [tasty]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[status]</td>
<td>Casado [married]</td>
<td></td>
</tr>
<tr>
<td>Underlying Structure</td>
<td>[unidirectional process]</td>
<td>Juan es alto [Juan is tall]</td>
<td>Does the adjective possess an underlying dynamic process? Is it unidirectional?</td>
</tr>
<tr>
<td></td>
<td>[bidirectional process]</td>
<td>Juan está mojado [Juan is wet]</td>
<td>Is it bidirectional?</td>
</tr>
<tr>
<td></td>
<td>[unidirectional event]</td>
<td>Juan es soltero [Juan is single]</td>
<td>Does the adjective possess an underlying dynamic event? Is it unidirectional?</td>
</tr>
<tr>
<td></td>
<td>[bidirectional event]</td>
<td>Juan está levantado [Juan is up]</td>
<td>Is it bidirectional?</td>
</tr>
<tr>
<td></td>
<td>[none]</td>
<td>Juan es inteligente [Juan is intelligent]</td>
<td>No underlying process or event</td>
</tr>
<tr>
<td>Gradiency</td>
<td>[+ gradient]</td>
<td>Juan está mojado [Juan is wet]</td>
<td>Can the adjective be put in a &quot;more or less&quot; construction?</td>
</tr>
<tr>
<td></td>
<td>[- gradient]</td>
<td>El ser humano es mortal [The human being is mortal]</td>
<td></td>
</tr>
</tbody>
</table>
The group of variables in Table 4.3 includes variables related to the adjective. These variables include *resultant state* that answers the question of whether the adjective is a resultant state (i.e., a verbal derivative that underlies a process or an event that leads to a state) or whether it is not. For example, *adoptado* ‘adopted’ is a *resultant state* derived from the verb *adoptar* “to adopt”. *Adjective class* refers to what semantic class best describes the adjective. Because I did not include *animacy* as a variable, all instances of physical characteristics in this corpus were categorized as *description of a person(ality)*. From Clements’ (2006) theoretical approach to the study of *copula+adjective* construction and Aguilar-Sanchez’ (2007) pilot study, two new linguistic independent variables were included. The first one is *underlying structure* which refers to whether the adjective has an underlying dynamic situation and what type of situation it is (i.e., unidirectional process or event and/or bidirectional process or event, or none). For example, in the sentence *Juan es alto* ‘Juan is tall’, the adjective *alto* ‘tall’ is said to have an underlying unidirectional process because a person can only grow tall. In the sentence *Juan está mojado* ‘Juan is wet’, the adjective *mojado* ‘wet’ is said to have an underlying bidirectional process because one can get wet, dry out and vice versa. In the sentence *Juan es soltero* ‘Juan is single’, the adjective *soltero* ‘single’ is said to have an underlying unidirectional process because a person can only go from single to married, but not vice versa. In the sentence *Juan está levantado* ‘Juan is up’, the adjectives is said to have a bidirectional event because a person can get up and lie down and vice versa. In the sentence *Juan es inteligente* ‘Juan is intelligent’, the adjective is said not to have an underlying event or process. The other variable included was *gradiency* (Givón, 2001), which refers to whether the adjective is gradient or not. In the sentence *Juan está mojado*
‘Juan is wet’ the adjective is said to be *gradient* because it can be found in a “more or less” construction as in more o less wet; whereas in the sentence *El ser humano es mortal* ‘The human being is mortal’, the adjective is said to be *absolute* or *not gradient* because *mortal* cannot be found in such construction.

Table 4.4

Coding of Independent Social Variables for the Study of Copula Choice

<table>
<thead>
<tr>
<th>Variable</th>
<th>Categories</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>[1] male</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2] female</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>16-91 continuous variable</td>
<td></td>
</tr>
<tr>
<td>Level of formal education</td>
<td>[1] Elementary incomplete</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2] Elementary complete</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[3] Junior High complete</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[4] High school complete</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[5] University</td>
<td></td>
</tr>
<tr>
<td>Bilingualism</td>
<td>[1] Monolingual</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2] Bilingual</td>
<td></td>
</tr>
<tr>
<td>Level of contact</td>
<td>[1] Restricted Contact</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2] Non-restricted Contact</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4 presents the rest of the independent variables included in the present study of a non-linguistic nature. The variables include *gender* which refers to the biological gender of the participant: male or female. *Age* is taken as a continuous variable in order to avoid polychotomization or the arrangement of continuous variables into three or more categories. *Level of formal education* is set as one of the variables due to the fact that minimum wages are government regulated (Inman, et al., 1997). *Profession* is divided into *services, home, business, education, farming, retired, and student*. *Bilingualism* separates bilinguals from monolinguals. This variable was composed taking
into account four questions from the background questionnaire. These questions are related to age when the speaker started learning a language, what languages the speaker reported speaking, whether the speaker reports being literate in such languages and his/her report of comfort speaking each of the languages. Numerical values from zero to 3 were assigned to each category. Then, a composite score was derived from the sum of those individual scores. Three different scores resulted. A score of 5 was assigned to the category monolingual (i.e., people reporting on speaking, reading, and writing Spanish), a score of 7 and up being reported bilingual in two languages (i.e, speakers reporting speaking English and Spanish and literacy in one or both languages), and 11 and up being multilingual (i.e., speakers reporting speaking more than two languages and being literate in one, two or all the languages).

Level of contact was also composed on a numerical scale created from each category found in two questions. These questions asked who the participant speaks each language with the most, and where s/he speaks each language the most. Each category of language use was restricted to only one group or place and was assigned a score of one and the others a score of two. Zero was assigned when the speaker reports not speaking one of the languages. Two sets of scores were devised. The first set is related to each language individually. A score of two is assigned to instances where the speaker reports speaking the language to everyone and a score of one when the speaker reports a restricted group of people s/he speaks that language to (e.g., only with relatives or only with friends). The same procedure was conducted for the place where the speaker reports speaking each language (e.g., only at home or everywhere). Total Scores for both languages were added and values of 5 and below were assigned a category of
restricted/indirect contact (i.e., speakers who reported speaking both languages in
different restricted areas) and scores of 6 and above were assigned to a category of not
restricted/direct contact (i.e., speakers reporting the use of both languages everywhere).

Attitude refers to whether or not the participant believes that it is important to
preserve the indigenous languages of Limón such as Limonese English.

Data Coding Procedure

In order to code data for the present study, the following procedure was followed.
The first step was to assign a code to each participant and its information. Each interview
was matched with its corresponding background data by means of this unique code. This
code was of the form <M32B03IVA>.. The letter M stands for the gender of the speaker,
in this case male. The two numbers following it indicate the participant’s age. The second
letter and two numbers indicate the recording folder and recording number in the Sony
ICD-U70 where the recording was stored, in this case folder B recording 31. The last
three letters indicate who helped the researcher that day with the interviews (i.e. contacts
in the community).

Each interview was matched with its corresponding background questionnaire.
Information in the background questionnaire was entered into an Access database built
for the purposes of facilitating the retrieval of information during the analysis process.
This database included all information in the background questionnaire and the recording
time of each interview and assigned a number to each speaker from 1-58.

Once the database was complete, the researcher proceeded to the transcription of
each interview. Two research assistants assisted with the transcription of the interviews.
The researcher checked each transcription carefully to avoid mistakes in them. The code
used to transcribe was a simplified code (See Table 4.5). Because the aim of this
dissertation was the collection of instances of *ser* and *estar*, transcription of sounds were not coded for unless it was very obvious that the person did not say the word in its entirety (e.g., toins instead of entonces ‘then’) (See Appendix 3 for a sample transcription).

Table 4.5

Transcription Code

<table>
<thead>
<tr>
<th>CODE</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>I:</td>
<td>Interviewer</td>
</tr>
<tr>
<td>S:</td>
<td>Subject/Participant</td>
</tr>
<tr>
<td>I2:</td>
<td>Second Interviewer</td>
</tr>
<tr>
<td>S2:</td>
<td>Other person</td>
</tr>
<tr>
<td>//</td>
<td>Silence or pause</td>
</tr>
<tr>
<td>[XXXX]</td>
<td>untranscribable words</td>
</tr>
<tr>
<td>[…]</td>
<td>long pause</td>
</tr>
<tr>
<td>&quot;beach&quot;</td>
<td>word in other language</td>
</tr>
<tr>
<td>[Risas]</td>
<td>Laughter</td>
</tr>
<tr>
<td>[Ruido]</td>
<td>Noise that interferes with the transcription</td>
</tr>
</tbody>
</table>

Following the transcription and revision of the interviews, all transcriptions were edited for consistency in format and saved in a document. Each line was assigned a number in order to facilitate reference to the tokens in the text. In order to facilitate the finding of instances of *ser* and *estar* in the transcriptions, each instance of all forms of *ser* and *estar* in all tenses and aspects were searched and replaced for small capitals and underlined for **SER** and small capitals and double underlined for **ESTAR**.

The researcher used an IRISPen Executive 6 to scan all sentences where *ser* and *estar* were used to an MS Excel worksheet. Once all accounts of *ser* and *estar* were scanned and entered into an MS Excel sheet, they were tabulated according to the
element that follows them. These elements are noun phrases, prepositional phrases (locatives), past participles, adjectives, present participles, noun clauses, adjective clauses, and adverbial clauses. This procedure was done to account for the frequency distribution of *ser* and *estar + adjective* against other types of predicates. Once descriptive statistics were tabulated, each token of *ser* and *estar + adjective* was entered into an MS Excel spreadsheet that identified the line number where it could be found as well as the speaker number. Each token was codified according to the independent variables defined for this study. (See Tables 4.2, 4.3, and 4.4 above).

In order to test the reliability of the coding, a random sample of about 5% of the total number of observations (n=80) was extracted to give to two independent coders. Coders were trained how to code tokens with a sample of ten tokens extracted from the corpus. They were given an MS Access database form with questions to answer according to the sentence they saw in a different panel. These questions captured the operationalization of each independent variable like those in Table 4.2. The independent coders were asked to code each observation extracted from the corpus. The use of two independent coders provided evidence of reliability and validity of my coding. All three datasets were compared to calculate the correlation coefficient as evidence of reliability (Cronbach’s alpha). Any discrepancy among the tokens was resolved by taking the coding given by two of the coders as the final code. If all three disagree, the researcher coded the token one more time in order to break the tie. Table 4.6 shows the results of a reliability analysis conducted in SPSS 16 (Statistical Package for the Social Sciences).
Table 4.6
Evidence of Reliability of the Code: 3 Coders

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cronbach's alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicate Reading</td>
<td>0.741</td>
</tr>
<tr>
<td>Susceptibility to change</td>
<td>0.822</td>
</tr>
<tr>
<td>Frame of reference</td>
<td>0.771</td>
</tr>
<tr>
<td>Experience with the referent</td>
<td>0.578</td>
</tr>
<tr>
<td>Adverb</td>
<td>0.930</td>
</tr>
<tr>
<td>Subject</td>
<td>0.965</td>
</tr>
<tr>
<td>Adjective class</td>
<td>0.856</td>
</tr>
<tr>
<td>Underlying structure</td>
<td>0.894</td>
</tr>
<tr>
<td>Gradiency</td>
<td>0.907</td>
</tr>
</tbody>
</table>

As shown in Table 4.6, the reliability coefficients range from moderate (.478) to the desired .8 or above. Because the Cronbach’s alphas were within this range, my coding of each token in this corpus is said to be reliable.

The statistical test and Data analyses

As explained above, the purpose of this dissertation is to predict how copulas are chosen in the context of *copula + adjective*. As argued in the Chapter 3 section on the statistic test, the appropriate statistical analysis for the prediction of copula choice is the logistic regression. In sociolinguistics a modified version of this analysis is used to conduct variable rule analyses (Tagliamonte, 2006). However, the fact that my data structure is hierarchical, leads to the violation of the assumptions of such a test because logistic regression assumes that the data comes in a non-hierarchical/not grouped manner; therefore, a multivariate logistic regression is not appropriate to analyze data of this type. Second, I want to avoid what Hox (2002) calls the “Simpson’s Paradox” which states that the problem of drawing completely erroneous conclusions may arise if grouped data,
drawn from heterogeneous populations, are collapsed and analyzed as if they came from a single homogeneous population. In order to meet all the assumptions made by the statistical test and to avoid drawing completely erroneous conclusions, I have chosen to use a multilevel logistic regression for the analysis of data in the present study. In order to conduct such an analysis, I used R (R Development Core Team, 2009), a statistical package for the behavioral and social sciences. As stated in Chapter 3, the statistical software R is an-open source software and is widely used for advanced statistical analysis due to its flexibility for modeling. The function “lmer()” stands for “linear mixed effects in R” in the “arm” package (Gelman, Su, Yajima, Hill, Gracia Pittau, Kerman & Zheng, 2009) which allowed the analysis of both levels of these data at the same time.

**Hypothesis testing**

Proponents of this type of analysis state that it is a statistical error to summarize comparisons by statistical significance and to draw a sharp distinction between significant and nonsignificant results (Gelman & Hill, 2007; Hox, 2002). Others arguing against the over-reliance on significant-nonsignificant dichotomy clarify that a statistically significant outcome only indicates that it is likely that there is some relationship between the variables (Vaske, 2002) and not that a $p$-value of .01 is a better than a $p$-value of .05. Hox (2002) states that the approach of summarizing by statistical significance has two drawbacks. The first, he states, is that statistical significance does not equal practical significance. The second is that changes in statistical significance are not themselves significant because only a small change is required to move from significant to nonsignificant.

By using this method of analysis, I am departing from the tradition of reporting results only by their significance coefficients. I will interpret the results following the
standards set forth by proponents of the multilevel analysis. I will still report significance coefficients when appropriate to support further inference. What follows is an example of the interpretation of a model fit within the multilevel analysis.

**Interpretation of the results**

For illustration purposes, I present the model *underlying structure 02* (a model from the selection of model process presented in Chapter 5). This model includes the predictors at level one as random effects: *underlying structure* and *gradiency*. At Level 2 as fixed effects, it contains the social predictors *age*, *education*, *bilingualism*, *contact*, *attitude*, and *gender*. The “lmer()” function works for generalized linear models (Gelman and Hill, 2007) and yields results in two parts in the software output. Table 4.7 presents these results. In the first column, we find the level of the analysis. Level 1, in this case, contains results of the role of linguistic predictors. Level 2 contain results of the role of social predictors. The parameter and the variables are the names of the parts of the model. Remember that the intercept is the parameter obtained when all the predictors are equal to zero. Unfortunately, the intercept is not easily interpretable because we are dealing with categorical variables and some of the predictors do not have zero as one of their categories. Therefore, the intercept becomes of no interest in this study.
Table 4.7  
Example of "lmer()" outcome in R:  
Model Fitted with Two Level-1 Predictors

<table>
<thead>
<tr>
<th>Level</th>
<th>Parameter</th>
<th>Variable</th>
<th>Coef.Est</th>
<th>Coef.SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>(Intercept)</td>
<td></td>
<td>-2.06</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Underlying Structure</td>
<td></td>
<td>0.66</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Gradiency</td>
<td></td>
<td>0.26</td>
<td>0.15</td>
</tr>
<tr>
<td>02</td>
<td>(Intercept)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td></td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td></td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td></td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bilingualism</td>
<td></td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contact</td>
<td></td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Save a Language</td>
<td></td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Residual SD</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td></td>
<td>1494.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DIC</td>
<td></td>
<td>1476.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deviance</td>
<td></td>
<td>1476.30</td>
<td></td>
</tr>
</tbody>
</table>

The slope is the predicted difference in the logit curve by each unit in that category of predictor. The parameter estimates that are presented are the coefficient estimate which means the predicted value for that predictor and the coefficient standard error of that predictor at level 1 and the standard deviation at level 2.  

Up to this point, at level 1, the procedures to arrive to these results are the same one used by GoldVarb with the slight difference that social/grouping predictors are included at level 2 and not aggregated as several instances of the same predictor. Each
coefficient estimate can be divided by 4 to achieve the difference between categories of each predictor. This is an approximation of the probabilistic weight given by GoldVarb because they tell how to interpret the influence of each category within a predictor group on the probability that *estar* is present. A factor is statistically significant in the model if its coefficient estimate is two standard errors above zero (the mean).

So, for the example in Table 4.7, we can infer that the variable underlying structure is a significant predictor because 0.05 * 2 = 0.10 and the coefficient estimate is 0.66. On the contrary, the predictor *gradiency* is not statistically significant because 0.15 * 2 = 0.30 and the coefficient estimate is only 0.24. In Chapter 6, statistically significant predictors are singled out with a single asterisk (*) next to the coefficient standard error.

The difference between categories is derived by dividing the coefficient standard error by 4. This procedure is a mathematical function that yields the maximum difference in the probability corresponding to a unit difference among predictor categories. For a full account of this procedure see Gelman and Hill’s (2007) Chapter 5. Because differences are presented in the probability scale, the best way to show these differences is in a graph.

In logistic regression, the probability logit curve of finding *estar* should look like an “S” shape which is steepest where the predictor value and the probability scale is 0.5. This means that there are equal probabilities between the appearance of *estar* and *ser* (i.e., no difference).
Figure 4.3. Visual explanation of the outcome for the predictor *underlying structure*

Figure 4.4. Visual explanation of the outcome for the predictor *gradiency*. 
Figure 4.3 shows that even though, *underlying structure* is statistically significant at p=0.05, it helps the prediction of *estar* but the differences among categories are small (i.e., 0.05/4= 0.0125) or less than 2 percent difference in the probability scale. That is the reason why each line is almost horizontal. Figure 4.4 shows that even though *gradiency* is not statistically significant the predictions of *gradiency* approximate the “S” shape, the differences are still too small (0.15/4= 0.0375) or about 3% of difference in the probability scale. As a result, we can infer that this model is not powerful enough to explain appearance of *estar*.

Social variables are analyzed in a similar fashion. Their standard deviation divided by four yields the difference over and above the differences explained by predictors at level 1 (i.e., linguistic factors). The coefficient that results from this division is translated as the average percentage of difference among categories. If the standard deviation equals zero, the groups do not differ; therefore, they do not contribute to the prediction of *estar*.

Gelman and Hill (2007) propose that it does not make sense to calculate residual standard deviation and $R^2$ for logistic regression and other discrete-data models because the models are not fit by least squares. Although there are pseudo $R^2$ values for logistic regression, what is standard, on the other hand, is to use the deviance statistic. The deviance statistic is a summary of model fit defined for logistic regression and other generalized linear models and is analogous to the residual standard deviation (Gelman & Hill, 2007; Hosmer & Lemeshow, 1989; Hox, 2002). According to Gelman and Hill (2007), the deviance statistic has three properties. The first one is that it is a measure of error and a lower deviance means a better fit to the data. The second is that if a predictor
that is simply random noise is added to a model, the deviance is expected to decrease by 1 on average. Third, when an informative predictor is added to a model the deviance is expected to decrease by more than 1. Gelman and Hill (2007) state that when \( k \) predictors are added to a model, deviance is expected to decrease by more than \( k \).

For example, let us compare \textit{underlying structure 02} model (model names from the selection of model process presented in Chapter 5) with another model. Let us call this new model \textit{underlying structure 01}, which includes only the variable \textit{underlying structure} with a deviance statistic of 1479.50. Underlying structure 02 has a deviance of 1476.3. The difference between these two deviances is 3.20. This difference is evaluated against the critical value of a \( \chi^2 \) with two degrees of freedom at \( p=0.05 \). The critical value is 5.99. The difference in deviances is smaller than this critical value; therefore, adding the variable gradiency does not increase the fit of the model a statistically significant amount. As a result, we should keep the model without \textit{gradiency}.

This analysis is comparable to that performed by GoldVarb because GoldVarb uses comparison of the loglikelihood coefficients to carry out the selection of models during the up and down run (Tagliamonte, 2006). The only difference is that in the analysis proposed for this dissertation, the variance explained by the social variables is kept at the level they came from (i.e., Level 2) in order to avoid violating the assumptions of logistic regression analysis. This separation between social variables and linguistic variables, to my knowledge and as evident from the review of the previous literature, has not been done in sociolinguistic studies using logistic regression.

\textbf{Selecting the variables}

For the purposes of this dissertation all predictors at level 1 were entered as fixed effects and predictors at level 2 were entered as random effects. The selection of the
variables for further analysis beyond this point was based on Gelman and Hill’s (2007) general principles for the inclusion of variables for further modeling.

These principles are summarized as follows:

1. Include all predictors that, for substantive reasons, might be expected to be important in predicting the outcome.
2. It is not always necessary to include these inputs as separate predictors; we can add them up and include the totals.
3. For predictors that have large effects, include their interactions as well.
4. Strategy for decisions regarding whether to include or exclude a predictive variable based on their sign and statistical significance:
   a. If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help dramatically, but it will not hurt them either.
   b. If a predictor is not statistically significant and does not have the expected sign, remove it from the model.
   c. If a predictor is statistically significant and does not have the expected sign, include it if it makes sense theoretically.
   d. If a predictor is statistically significant and has the expected sign, include it.

By following these principles, I depart from the tradition of relying only on the significance level to include or exclude variables in a model. This departure from significance testing is one of the operating principles of multilevel modeling because what we are in search of is how to explain the influence of a social factor in a linguistic phenomenon. Therefore, I included predictors that theoretically made sense.

Each variable was entered into a multilevel logistic regression and analyzed according to its significance and the change it produced in the deviance statistic. Variables that represented a significant drop in the deviance of each model regardless of their statistical significance were selected for the creation of possible models. Based on the results from the preliminary logistic regression on each predictor by itself, all possible combinations of these variables were calculated and they are regarded henceforth as
models. Because this is an exploratory stage and the fact that a smaller deviance means a better fit, deviances for each group were compared and the lowest deviance was selected from each group.

Following this first selection, all selected models were compared to determine what model best fits the data. Both levels of the data were taken into account and generalizations are made based on the results. Tables and charts are used to provide an easy reading of the analyses and the data. All analyses relevant to the results are presented in Chapter 5 in a detailed manner in order to allow the replication of this study. Any discussion of the results and how they fit the bigger picture are reserved for Chapter 6.
Chapter 5

Results

Introduction

The purpose of this chapter is to present the findings of the analysis of *copula + adjective* in the Spanish of Limón, Costa Rica. Following the results section, I present a discussion that explores which linguistic variables predict the use of *ser* and *estar* in the context of *copula + adjective* in the Spanish spoken by Costa Ricans in Limón as well as the role of social factors in the prediction of such variables. Finally, this section also explores the degree of stability of the phenomenon under study. As I argued in Chapter 2, the use of designs that take into account multiple variables for the study of copula choice have provided important information regarding the prediction of features that help us understand copula choice behavior in different linguistic and social settings. I also argued that in any new research project in search of an explanation for copula choice in Spanish, all different factors should come from well-founded theoretical accounts and experiments. I include syntactic, semantic and pragmatic factors in the analysis of copula choice because including multiple factors provides us with a better outcome. Key methodological issues were addressed, as presented in Chapter 3, in order to achieve the desired levels of generalization needed to claim a complete or near complete account on this phenomenon. The results presented in this chapter represent the foundation for the discussion carried out in Chapter 6.

The chapter is divided into three sections. The first one is devoted to the presentation of the descriptive statistics and the general tendencies found in the data. Second is the presentation of the statistical procedures performed to find a model that is
an appropriate fit to the study of copula choice in this sample. I find justification to present this section in this chapter in the fact that in order to arrive at the desired statistical model, I performed several statistical calculations. Some of these processes are done automatically by software such as GoldVarb and SPSS. Because one of the goals of this dissertation is methodological in nature, I have decided to present the results of my calculations together with those pertaining to prediction of copula choice. By presenting the calculations, I intend to emphasize the need to know the data structure very well and the need to depart from an over-reliance on significance testing. The third is the presentation of the analysis of the selected model and the results of such a model for the study of copula + adjective in the Limonese Spanish.

Descriptive statistics

The sample

A total of twenty hours and twenty five minutes of recorded speech from 58 participants were collected and transcribed for the present study. From this transcription a total of 5795 instances of ser and estar were found. From these 5,795, only 1,632 were instances of ser and/or estar + adjectives; therefore, they were extracted for analysis. Although unrelated to my research questions, the distribution of the rest of the tokens is presented here to exemplify the extent of the phenomenon in the copular system of the Spanish of Limon.
Table 5.1

Distribution of *ser* and *estar* according to the type of context

<table>
<thead>
<tr>
<th>Complement</th>
<th>Ser</th>
<th>Estar</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>1218</td>
<td>236</td>
<td>1454</td>
</tr>
<tr>
<td>Past Participle</td>
<td>66</td>
<td>112</td>
<td>178</td>
</tr>
<tr>
<td>Adjective Clause</td>
<td>87</td>
<td>9</td>
<td>96</td>
</tr>
<tr>
<td>Noun (Phrase)</td>
<td>2074</td>
<td>167</td>
<td>2241</td>
</tr>
<tr>
<td>Prepositional Phrase</td>
<td>319</td>
<td>226</td>
<td>545</td>
</tr>
<tr>
<td>Present Participle</td>
<td>12</td>
<td>393</td>
<td>405</td>
</tr>
<tr>
<td>Infinitive Clause</td>
<td>80</td>
<td>1</td>
<td>81</td>
</tr>
<tr>
<td>Sentence or Adverbial Clause</td>
<td>229</td>
<td>62</td>
<td>291</td>
</tr>
<tr>
<td>No complement (e.g., false starts)</td>
<td>192</td>
<td>68</td>
<td>260</td>
</tr>
<tr>
<td>Discursive Phrase</td>
<td>244</td>
<td>0</td>
<td>244</td>
</tr>
<tr>
<td>Total</td>
<td>4521</td>
<td>1274</td>
<td>5795</td>
</tr>
</tbody>
</table>

Table 5.1 presents the distribution of each instance of *ser* and *estar* by the type of complement that accompanied the copula. Each instance of *ser* and *estar* was classified according to its complement, whether they were unfinished ideas (i.e., no complement), or whether they were discursive phrases such as “es más” *besides* or “es que” *ehem*. The first column on Table 5.1 contains the type of complement. The second contains the count of each category by copula. The last one contains the total instances in that category. The total of 1,632 comes from the categories *adjective* and *past participle*. *Past participles* were chosen here for their dual function as *past participles* and as *adjectives*.

As stated in Chapter 3, instances of *ser/estar + adjective* account for roughly thirty percent of the total number of instances of *ser* and *estar* in this population. It is important to report this number here because it will help with the design of future studies especially in the determination of appropriate sample sizes. As I argued in Chapter 3, the
magnitude of the presence of the phenomenon in a system is important in calculating sample sizes because it allows the researcher to use real population characteristics and not suppositions.

**General Tendencies**

All variables were coded according to the coding scheme outlined in Chapter 4. What follows are the descriptive statistics of each variable. Because all but one variable are categorical, only the totals per category of the variable and the copula are presented. For the variable *age*, I present the mean, median and the range. These three statistics are important for the analysis and discussion of this variable because they tell us basic information regarding its distribution. As I mentioned in Chapter 4, the dependent variable was the presence of *ser* and *estar* in a structure of *ser/estar + adjective*. A total of 1284 instances of *ser* and 348 of *estar* were found. For the purposes of codification as outlined in Chapter 4, variables were divided into two categories. The first category contains variables related to the predicate (i.e., the verb and its complements). The second contains features related to the adjective only.

In this section, I present a description of the general tendencies found in this data set. Each variable is presented separately in its own table and they are accompanied by a figure with graphical information on that variable. I present general tendencies here as a preliminary look at the data (Tagliamonte, 2006); however, I follow Tagliamonte’s (2006) word of caution that

“a distributional, factor by factor analysis […] offers numerous insights into the factors conditioning the occurrence of your linguistic variable. However, tabulations of effects taken one at the time cannot reveal the combined impact of all the factors together. This is where variable rule analysis comes in as the right tool for the job. It can model the simultaneous operation of factor effects, reveal the relative importance of each one to the other, and select which ones are significant.” (p. 215)
and these preliminary results are taken as just that, preliminary. All inferences of the
effect of the factors on the appearance of *estar* are drawn from the multilevel logistic
regression. Each table presents the data summarizing each variable and subcategories and
their distributions between *ser* and *estar*. Column one contains the variable and category
name. Column two contains labels for descriptive statistics. Column three contains the
count of each instance of *ser* and column four of *estar*. The last column shows the total
by category and percentage total of the variable. Any discussion related to these
tendencies and how they fit into the bigger picture is given in Chapter 6. For the purposes
of this chapter, I present only statistical results.

**Linguistic variables**

Linguistic variables

Table 5.2 and Figure 5.1 show that *estar* is used only with predicates with a *stage-
level reading* (74%).

Table 5.2

Distribution of *ser* and *estar* according to the predictor *predicate reading*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Subcategory</td>
</tr>
<tr>
<td>Individual-level</td>
<td>n</td>
<td>1160</td>
<td>0</td>
<td>1160</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>100%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Stage-level</td>
<td>n</td>
<td>124</td>
<td>348</td>
<td>472</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>26%</td>
<td>74%</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
</tr>
</tbody>
</table>
Individual-level reading, as expected, has coupled only with *ser* as pointed out by Clements (2006) and the Stage-level-Individual-level hypothesis (Kratzer, et al., 1995).

![Figure 5.1. Distribution of *ser* and *estar* according to predictor predicate reading.](image)

Table 5.3 and Figure 5.2 present the distribution of the variable susceptibility to change.

Table 5.3

Distribution of *ser* and *estar* according to the predictor susceptibility to change

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Subcategory</td>
</tr>
<tr>
<td>(+) Susceptible</td>
<td>n</td>
<td>1104</td>
<td>337</td>
<td>1441</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>77%</td>
<td>23%</td>
<td></td>
</tr>
<tr>
<td>(-) Susceptible</td>
<td>n</td>
<td>180</td>
<td>11</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>94%</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
</tr>
</tbody>
</table>
Figure 5.2. Distribution of *ser* and *estar* according to the predictor *susceptibility to change*

*Estar* appears more with adjectives codified as *susceptible* to change (23%) than with those that were *not susceptible* to change (6%).

Table 5.4 and Figure 5.3 show the tendencies found for the predictor group *experience with the referent*.

**Table 5.4**

Distribution of *ser* and *estar* according to the predictor *experience with the referent*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
<th>Variable % Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Subcategory</td>
<td></td>
</tr>
<tr>
<td>Ongoing</td>
<td>n</td>
<td>1241</td>
<td>185</td>
<td>1426</td>
<td>87%</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>87%</td>
<td>13%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immediate</td>
<td>n</td>
<td>36</td>
<td>160</td>
<td>196</td>
<td>12%</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>18%</td>
<td>82%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirect</td>
<td>n</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>70%</td>
<td>30%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
<td>100%</td>
</tr>
</tbody>
</table>
Figure 5.3. Distribution of *ser* and *estar* according to the predictor *experience with the referent*.

*Estar* is found more with predicates that were coded as the speaker having an *immediate experience with the referent* (82%). It was found with less frequency with those coded as the speaker having an *indirect experience with the referent* (30%). And with predicates coded as having an *ongoing experience with the referent* with the lower frequency (13%).

Table 5.5 and Figure 5.4 present the distribution of the predictor *adverb*. This variable shows an interesting pattern.
Table 5.5

Distribution of *ser* and *estar* according to the predictor *adverb*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Subcategory</td>
</tr>
<tr>
<td>Overt</td>
<td>n</td>
<td>295</td>
<td>111</td>
<td>406</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>73%</td>
<td>27%</td>
<td></td>
</tr>
<tr>
<td>Covert</td>
<td>n</td>
<td>359</td>
<td>194</td>
<td>553</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>65%</td>
<td>35%</td>
<td></td>
</tr>
<tr>
<td>No adverb</td>
<td>n</td>
<td>630</td>
<td>43</td>
<td>673</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>94%</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
</tr>
</tbody>
</table>

![Distribution of ser and estar according to the predictor adverb](image)

Figure 5.4. Distribution of *ser* and *estar* according to the predictor *adverb*

*Estar* appears more with adverbs that are *covert* or *overt* (35% and 27%) respectively. I will address this pattern in Chapter 6 because it contradicts results in previous empirical accounts.

Table 5.6 and Figure 5.5 present the distribution of the last variable related to the sentence: *subject.*
Table 5.6

Distribution of *ser* and *estar* according to the predictor *subject*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>%</td>
<td>% Total</td>
</tr>
<tr>
<td>First Order</td>
<td></td>
<td>785</td>
<td>75%</td>
<td>266</td>
</tr>
<tr>
<td>Second Order</td>
<td></td>
<td>499</td>
<td>86%</td>
<td>82</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>86%</td>
<td>348</td>
</tr>
</tbody>
</table>

Figure 5.5. Distribution of *ser* and *estar* according to the predictor *subject*

*Estar* appears with lower frequency with second-order entities (14%) and with higher frequency with first order entities (25%).
The adjective

The next groups of predictors are related to the adjective itself. Table 5.7 and Figure 5.6 display the distribution of the variable resultant state.

Table 5.7

Distribution of *ser* and *estar* according to the predictor resultant state

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subcategory</td>
<td></td>
<td></td>
<td>Variable</td>
</tr>
<tr>
<td>(-) Resultant</td>
<td>n</td>
<td>1159</td>
<td>176</td>
<td>1335</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>87%</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>(+) Resultant</td>
<td>n</td>
<td>125</td>
<td>172</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>42%</td>
<td>58%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
</tr>
</tbody>
</table>

Figure 5.6. Distribution of *ser* and *estar* according to the predictor resultant state
As we can see, *estar* appears more with adjectives that are coded as being a *resultant state* (58%) and less with adjectives that were coded as *non-resultant states* (13%).

Table 5.8 and Figure 5.7 depict the distribution of the predictor *adjective class*.

### Table 5.8

#### Distribution of *ser* and *estar* according to the predictor *adjective class*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Variable</td>
<td>%</td>
<td></td>
<td>Subcategory</td>
</tr>
<tr>
<td>Age</td>
<td>n</td>
<td>69%</td>
<td>31%</td>
<td>67</td>
</tr>
<tr>
<td>Size</td>
<td>n</td>
<td>96%</td>
<td>4%</td>
<td>47</td>
</tr>
<tr>
<td>Origin/Profession/Religion</td>
<td>n</td>
<td>100%</td>
<td>0%</td>
<td>68</td>
</tr>
<tr>
<td>Description/Evaluation</td>
<td>n</td>
<td>85%</td>
<td>15%</td>
<td>958</td>
</tr>
<tr>
<td>Description of Person(ality)</td>
<td>n</td>
<td>94%</td>
<td>6%</td>
<td>172</td>
</tr>
<tr>
<td>Color</td>
<td>n</td>
<td>100%</td>
<td>0%</td>
<td>74</td>
</tr>
<tr>
<td>Mental/Physical State</td>
<td>n</td>
<td>14%</td>
<td>86%</td>
<td>125</td>
</tr>
<tr>
<td>Sensory characteristic</td>
<td>n</td>
<td>86%</td>
<td>14%</td>
<td>21</td>
</tr>
<tr>
<td>Status</td>
<td>n</td>
<td>43%</td>
<td>57%</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
</tr>
</tbody>
</table>
Figure 5.7. Distribution of *ser* and *estar* according to the predictor *adjective class*

Two categories of the variable *adjective class* showed a categorical distribution with *ser*. These categories were *origin/profession/religion* and *color* such as *limonense* ‘limonese’ and *negra* ‘black’.

*Estar* is found with more frequency with adjectives of *mental/physical state* (86%) followed by adjectives of *status* (57%) and *age* (31%). It was also found with relatively lower frequency with adjectives of *description/evaluation* (15%), *sensory characteristics* (14%), description of a *person(ality)* 6%, and adjectives of *size* (2%).

Table 5.9 and Figure 5.8 present the distribution of the variable *underlying structure*. 
Table 5.9

Distribution of *ser* and *estar* according to the factor group *underlying structure*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Subcategory</td>
</tr>
<tr>
<td>No underlying structure</td>
<td>n</td>
<td>1087</td>
<td>154</td>
<td>1241</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>88%</td>
<td>12%</td>
<td></td>
</tr>
<tr>
<td>Unidirectional Process</td>
<td>n</td>
<td>74</td>
<td>58</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>56%</td>
<td>44%</td>
<td></td>
</tr>
<tr>
<td>Bidirectional Process</td>
<td>n</td>
<td>49</td>
<td>68</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>42%</td>
<td>58%</td>
<td></td>
</tr>
<tr>
<td>Unidirectional Event</td>
<td>n</td>
<td>28</td>
<td>26</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>52%</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td>Bidirectional Event</td>
<td>n</td>
<td>46</td>
<td>42</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>52%</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
</tr>
</tbody>
</table>

Figure 5.8. Distribution of *ser* and *estar* according to the factor group *underlying structure*
Estar is found with more frequency with adjectives coded as having a bidirectional process (58%), unidirectional events (48%), bidirectional events (48%), and unidirectional processes (44%). Adjectives with no underlying structure were found only 12% of the time with estar.

Table 5.10 and Figure 5.9 present the results of the predictor gradiency.

Table 5.10

Distribution of ser and estar according to the factor group gradiency

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptive</th>
<th>Ser</th>
<th>Estar</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>455</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>82%</td>
<td>18%</td>
</tr>
<tr>
<td></td>
<td>Gradient</td>
<td>n</td>
<td>829</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>77%</td>
<td>23%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1284</td>
<td>348</td>
<td>1632</td>
</tr>
</tbody>
</table>

Figure 5.9. Distribution of ser and estar according to the factor group gradiency
Estar is found with a slightly higher frequency with gradient adjectives (23%) than with absolute adjectives (18%).

It is worth noting here that a preliminary look at the distribution of the data according to categories among variables seems to be in accord with the predictions of previous theoretical accounts of copula choice in Spanish discussed in Chapter 2. I discuss these, however, after I conduct the multilevel logistic regression to determine how each predictor behaves when put together with other factors in a model (i.e., variable rule analysis).

**Social Predictors**

After analyzing each social variable for its theoretical importance in helping predict copula choice in this sample, I decided not to include the variable *profession* and include only *education* because they were highly correlated ($\chi^2 = 2008$, df=30, p=0.05). The formulation of models was done taking into account all remaining variables (linguistic and social). Because salaries in Costa Rica are regulated by the Government and they are based on education level, I decided to exclude *profession* from the statistical model to predict *ser* and *estar* in the context *ser/estar + adjective* because, essentially, they represent the same variable. All other variables were included.

**Selection of Variables**

What follows is the procedure to select linguistic predictors according to their prediction power in the context of multilevel logistic regression. For now, I only present justifications for decisions made during the analyses of the data that affected the creation of models and the results from those analyses. Theoretical foundations for this issue were presented in Chapter 4 while further discussion of this issue follows in Chapter 6.
Social Variables

Social variables are exempt from the logistic regression or the significance criterion as mentioned in Chapter 4. It is worth remembering that it is not appropriate to use significance as the only criterion for including particular group indicators in a multilevel model (Gelman & Hill, 2007). The justification for this reasoning, Gelman and Hill (2007) propose and I adhere to, is that we need to include all grouping indicators because nothing is lost in doing so. Furthermore, they state that the purpose of a multilevel model is not to see whether elements in grouping factors are different statistically from one another, but to study how grouping factors help explain the phenomenon under study. In sociolinguistic terms we do not seek to determine whether one education group is different from another, but how the variable education helps us predict the use of estar. Even though the tradition has been to create age groups to understand differences in the production of a variable, in the study of ser and estar this relationship has been very difficult to find. With a multilevel model, I do not seek to prove if one age is different from another, but to see if age explains some of the variance. Any prediction based on these variables is done when the model is fitted and the inference on their role is done based on the amount of variability found among groups. In other words, to study change, we look for the influence of age in the explanation of variance at Level 2. Then, we examine whether age presents a trend of change that can be attributed to the variance explained by the predictor age. Therefore, we are not looking to see if one age is different from the other. Even though I depart from the tradition of creating age groups and comparing them as explained in Chapter 3, the inferences made here seek to explain the same phenomenon: the influence of a social factor on a linguistic phenomenon.
Up to this point, I have conducted the first steps of a variable rule analysis. First, I had a look at the data to discover possible tendencies. The following procedures differ from the traditional variable rule analysis in that social predictors are included in the selection of linguistic variables.

**Linguistic Variables**

As stated above, in order to formulate models to compare and analyze for the present study, variables were tested for their prediction strength. I conducted individual multilevel logistic regressions for each potential linguistic predictor in order to decide whether or not to include it in the creation of models for the analysis. Although a Chi-square is an appropriate method to test the relationship of individual variables to the dependent variable and it has been used widely in studies of copula choice in Spanish (Geeslin, 2002a; Geeslin & Guijarro-Fuentes, 2008; Silva-Corvalan, 1994), conducting individual multilevel logistic regression on each variable allowed me to determine whether the model where the variable is introduced alone significantly explains variance in the data and not just whether the predictor group is associated with the outcome variable.\(^8\)

As stated in Chapter 4, by studying the drop in the deviance statistic from the null deviance (i.e., deviance of the model with no predictors: the intercept), we can determine whether the inclusion of the variable helps the prediction of *estar* regardless of its significance. Let us remember that the closer to zero de deviance statistic is, the better the fit to the data. This procedure goes beyond determining the level of association to the dependent variable and provides the researcher with accurate coefficients of its

---

\(^8\) Variance explains the behavior of the data and by looking for the best fit (i.e., less statistical error), we can explain the phenomenon under study from a statistical point of view.
predictability value to a model. Let us remember that a distributional analysis, the one presented above, does not necessarily have to agree with the rest of the analyses. This happens because predictors behave differently when they are introduced in a particular statistical model with other predictors. In this case, every predictor was introduced in a multilevel logistic regression with the social variables at level 2 (i.e., they help explain the variance not explained by the linguistic predictors alone). This procedure is innovative in the field of sociolinguistics. Therefore, results achieved from now on, could not be compared with those that might arise by using GoldVarb because the theoretical and mathematical assumptions are different at this point. However, the theoretical interpretations of the findings can be compared because we are studying the same phenomenon. Any differences in the final results are due to the statistical approach employed or to differences in the Spanish in Limón and not the notion of copula choice that we have constructed over the years.

Table 5.11

Predictive value of individual predictors

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Null Deviance</th>
<th>Residual Deviance</th>
<th>Difference</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicate Reading</td>
<td>1691.40</td>
<td>543.60</td>
<td>1147.80</td>
<td>X</td>
</tr>
<tr>
<td>Susceptibility to Change</td>
<td>1691.40</td>
<td>1651.70</td>
<td>39.80</td>
<td>X</td>
</tr>
<tr>
<td>Experience with the Referent</td>
<td>1691.40</td>
<td>1349.80</td>
<td>341.60</td>
<td>X</td>
</tr>
<tr>
<td>Adverb</td>
<td>1691.40</td>
<td>1534.40</td>
<td>341.60</td>
<td>X</td>
</tr>
<tr>
<td>Subject</td>
<td>1691.40</td>
<td>1662.10</td>
<td>29.40</td>
<td>X</td>
</tr>
<tr>
<td>Resultant State</td>
<td>1691.40</td>
<td>1445.20</td>
<td>246.30</td>
<td>X</td>
</tr>
<tr>
<td>Adjective Class</td>
<td>1691.40</td>
<td>1531.20</td>
<td>160.20</td>
<td>X</td>
</tr>
<tr>
<td>Underlying Structure</td>
<td>1691.40</td>
<td>1532.70</td>
<td>158.80</td>
<td>X</td>
</tr>
<tr>
<td>Gradiency</td>
<td>1691.40</td>
<td>1684.30</td>
<td>7.20</td>
<td>X</td>
</tr>
</tbody>
</table>

* Significance is based on a $\chi^2(p-.05, df 2) = 5.99$ for the drop in deviance
Table 5.11 presents the results of the analyses performed to check for the prediction power of each predictor when used by itself as a fixed effect with the social predictors at level 2 entered as random effects. As we can observe in Table 5.11, from all the variables included in the preliminary analysis of predictive strength with social variables at level 2, all variables, by themselves, were capable of explaining part of the variance found in the data even when a distributional analysis and patterns lead us to believe otherwise (i.e., susceptibility to change). I will return to this discussion in Chapter 6. Linguistic predictors included at level 1 were predicate reading, susceptibility to change, experience with the referent, adverb, subject, resultant state, adjective class, underlying structure, and gradiency.

The Models

Formulation of Models

There were a total of one hundred and twenty nine possible combinations of predictors. These combinations were designed by adding and removing variable combinations starting one variable at a time until the saturated model was reached. All models were grouped according to which predictor was introduced first. As a result, there were nine groups of models plus the saturated one (see Table 5.12).
Table 5.12

Number of Possible Models by First Variable Included

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of possible combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturated Model</td>
<td>1</td>
</tr>
<tr>
<td>Predicate Reading</td>
<td>36</td>
</tr>
<tr>
<td>Susceptibility to Change</td>
<td>29</td>
</tr>
<tr>
<td>Experience with the Referent</td>
<td>22</td>
</tr>
<tr>
<td>Adverb</td>
<td>16</td>
</tr>
<tr>
<td>Subject</td>
<td>11</td>
</tr>
<tr>
<td>Resultant State</td>
<td>7</td>
</tr>
<tr>
<td>Adjective Class</td>
<td>4</td>
</tr>
<tr>
<td>Underlying Structure</td>
<td>2</td>
</tr>
<tr>
<td>Gradiency</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>129</strong></td>
</tr>
</tbody>
</table>

Selection of the best fitting model

Table 5.13 presents the summary of the results from this selection and, for comparability reasons with other studies, predictors that were statistically significant in the prediction of estar.
Table 5.13

Model with the Lowest Deviance with their respective significant predictors

<table>
<thead>
<tr>
<th>Base Variable</th>
<th>Model No.</th>
<th>Deviance</th>
<th>Significant Predictors of the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturated Model</td>
<td>361.20</td>
<td>expref + adverb + subject + resultstate + adjclass</td>
<td></td>
</tr>
<tr>
<td>Predicate Reading</td>
<td>361.20</td>
<td>expref + adverb + subject + resultstate + adjclass suschang + expref + adverb + subject + resultstate + adjclass</td>
<td></td>
</tr>
<tr>
<td>Susceptibility</td>
<td>28</td>
<td>940.40</td>
<td>expref + adverb + subject + resultstate + adjclass</td>
</tr>
<tr>
<td>ExpeRefe</td>
<td>21</td>
<td>947.60</td>
<td>expref + adverb + subject + resultstate + adjclass</td>
</tr>
<tr>
<td>Adverb</td>
<td>16</td>
<td>1208.70</td>
<td>adverb + subject + resultstate + adjclass</td>
</tr>
<tr>
<td>Subject</td>
<td>11</td>
<td>1322.40</td>
<td>subject + resultstate + adjclass</td>
</tr>
<tr>
<td>ResultState</td>
<td>7</td>
<td>1328.70</td>
<td>resultstate + adjclass</td>
</tr>
<tr>
<td>AdjClass</td>
<td>4</td>
<td>1400.90</td>
<td>adjclass + unstruct</td>
</tr>
<tr>
<td>UnderStr</td>
<td>2</td>
<td>1476.30</td>
<td>unstruct</td>
</tr>
<tr>
<td>Gradien</td>
<td>1</td>
<td>1641.10</td>
<td>gradien</td>
</tr>
</tbody>
</table>

The first column contains the model group. The second column contains the model number within that group. The deviance column shows the deviance statistic for that model. The last column displays the variables that were ±2 standard errors from the mean (i.e., statistically significant at .05 level). The other predictors were not included in this table, but they are important in the construction of the statistical model. Those predictors are *predicate reading, susceptibility to change*, and *gradiency*.

At this point, I clarify that the use of statistical significance here is done to represent the possible difference among predictors, but not as a decision-making tool. Throughout this section and the next, I will use the statistical significance notation to allow for the comparison with other studies, but all decision-making is based on the predictability power of the model (i.e., the deviance). A deviance closer to zero represents a better fit to the data; therefore, a better model.
Even though some variables (e.g., gradiency or underlying structure) turn out to be significant in the multilevel logistic regression when included alone (e.g., Model 1 of the Underlying Structure group and Table 5.13 above); they are not significant in other models. Not only are they not significant, they do not fit the data as well as more complete models such as susceptibility to change Model 28, which is superior in fitting the data, because it possesses a much smaller deviance statistic\(^9\), but it still does not fit the data as well as predicate reading Model 36 or the saturated model. Predicate reading Model 36’s deviance is much smaller than the rest of the models. This comparison of differences in deviances means that the model with the biggest drop in deviance (i.e., with the deviance closer to zero) is Predicate reading Model 36’s which shows a better fit to the data.

Predicate reading Model 36 only differs from the saturated model in that the saturated model includes one more predictor: underlying structure. However, the inclusion of this predictor does not create a drop in the deviance; on the contrary, it remains the same. This fact renders predicate reading Model 36 as the best fit for the data. Thus, from this point forward all analysis of the data and results reported are those of predicate reading Model 36.

At this point all implications of using this model are purely statistical. Linguistic implications arise from the use of significance testing for interpreting the statistical model(s). If we made inference solely by looking at the predictors that are significant in a model with a large deviance, but the same predictors are non-significant in a model with a smaller deviance, we would arrive at the wrong conclusions. This is because a model with a smaller deviance has more statistical predictive power than a model with a

\(^9\) Deviance statistics need to be closer to zero
larger deviance. Thus, the factor that we may regard as significant is actually not significant. Additionally, if our model does not include enough predictors to explain statistical error in the data; we are using models with low predictive power, which lead us to the wrong conclusions.

*Predicate Reading Model 36*

As explained in Chapter 4, the multilevel analysis for this study has two levels. Level 1 contains all linguistic variables included in the model. Level 2 contains all social variables. Even though the tradition in the field is to look for the statistically significant predictors and explain outcomes based only on them, by using the multilevel approach I take statistically non-significant predictors and analyze them as essential in the prediction process. In other words, when using inferential methods to analyze linguistic data, these predictors contribute to the understanding of the phenomenon under study as part of the explanation of variance in the sample. This approach departs from our reliance on a statistical significance test to study sociolinguistic phenomena. It also departs from the notion of clear-cut differences or trying to find the only possible answer to our research questions and provides us with an approach that produces generalizable linguistic findings. This approach is based on our understanding of the phenomenon and is supported by the statistical predictions of the model. In the present work, this approach produced a more comprehensive approach for explaining copula choice in the context of *copula + adjective*.

*The multilevel logistic regression*

In this section, I present the results of a multilevel logistic regression analysis for the data I collected. Each of the predictors included in this model represent an integral
part of such a model even though they are not significant predictors of *estar*. As I present this section, each variable and its importance in the model is explained.

After the analysis and comparison of models, the best fit of the data was provided by *predicate reading* Model 36 which includes the predictors (1) *predicate type*, (2) *susceptibility to change*, (3) *experience with the referent*, (4) *adverb*, (5) *subject*, (6) *resultant state*, (7) *adjective class*, and (8) *gradiency* at the linguistic level (i.e., level 1) and the predictors (1) *age*, (2) *education*, (3) *bilingualism*, (4) *attitude*, (5) *contact*, and (6) *gender* at the social/grouping level (i.e., level 2).

In order to report the results produced by this model, I will present them in two subsections. The first section is related to the coefficients at level 1. In other words, the scores predicted by the model for each linguistic predictor included. These coefficients are the ones related to the token itself. They are the linguistic predictors of *estar*. The second section is related to the coefficients at level 2. These coefficients are related to the grouping variables. In other words, they are the social/speaker variables. The results and inferences are presented here, but the discussion of each is presented in Chapter 6.

Table 5.14 displays the outcome of the multilevel logistic regression analysis of *predicate reading* Model 36. This Table is divided vertically into four sections and horizontally into five sections including Level, Parameter, Predictor, Coefficient Estimates, Coefficient Standard Error, and whether the predictor is significant at .05. Vertically, section one displays level 1 components and statistics, section two displays level 2 components and statistics, section three displays counts, and section four displays model-related statistics such as the deviance. Vertical sections three and four are not
dealt with here because they have already been dealt with previously in this chapter. I concentrate here on sections one and two only.

**Linguistic predictors**

The first parameter that is calculated in a multilevel logistic regression at level 1 is the intercept. The intercept is the point at which all variables at level 1 are zero. Unfortunately, the intercept is not easily interpretable because we are dealing with categorical variables and some of the predictors do not have zero as one of their categories. Therefore, the intercept is of no interest in this study. In other words, the intercept does not have linguistic implications because of the characteristics of our predictors.
Table 5.14

Multilevel Logistic Regression of predicate reading Model 36

<table>
<thead>
<tr>
<th>Level</th>
<th>Parameter</th>
<th>Variable</th>
<th>Coef.Est</th>
<th>Coef.SE</th>
<th>p=0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 01</td>
<td>(Intercept)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Predicate Reading</td>
<td>22.41</td>
<td>84.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Susceptibility to Change</td>
<td>1.18</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experience with the Referent</td>
<td>1.95</td>
<td>0.34</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adverb</td>
<td>-0.61</td>
<td>0.24</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subject</td>
<td>-1.44</td>
<td>0.31</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Resultant State</td>
<td>2.30</td>
<td>0.44</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adjective Class</td>
<td>0.31</td>
<td>0.09</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gradiency</td>
<td>-0.19</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>Level 02</td>
<td>(Intercept)</td>
<td>Age</td>
<td>1.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Education</td>
<td>0.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bilingualism</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contact</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Save a Language</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gender</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Residual SD</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Observations</td>
<td></td>
<td></td>
<td>1632</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groups per Level 02 variables</td>
<td></td>
<td>Age</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gender</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Education</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bilingualism</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contact</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Save a Language</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td></td>
<td>391.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td></td>
<td></td>
<td>361.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deviance</td>
<td></td>
<td></td>
<td>361.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The second parameter or set of parameters calculated is the slope for each of the predictors, which includes the coefficient estimates of level 1 predictors and their respective standard errors. A predictor is significant if its coefficient estimate is ±2 standard errors from the mean. In this case the mean is zero. Each of these predictors is analyzed by its differences in the probability scale. The probability scale runs from zero being not probable and one being certainty of the event with 0.5 in the middle as equal probability. To understand this difference, we divide the coefficient estimates by 4 as explained in Chapter 4. The result is the difference in the probability scale. These results are similar to the probabilistic weights that software such GoldVarbX (Sankoff, Tagliamonte, & Smith, 2005) yields for which a weight above 0.5 means that the factor favors the probability that the dependent variable equals the application value. In this study, it represents the probability that the dependent variable selected is estar (i.e., estar=1). Thus, a positive difference means the group of parameters helps the probability of the appearance of estar by that percent in the probability scale. A negative value means that the group of parameters does not favor the probability of appearance of estar on the probability scale.

From Table 5.14 in column six we can see that the variable predicate reading is not a significant predictor of estar; however, as indicated by its coefficient estimate, it helps the model better explain the variance in the sample.

The next set of figures presented here can be read as follows:

a. The values on the axis $y$ are the probability scale and the values on $x$ are the category values of each predictor.
b. The inverse logit line is the regression line for each predictor or the transformation of linear predictors into probabilities employed in logistic regression.

c. The point where the predictor value and the regression line meet probability 0.5 is where there are equal probabilities of finding *estar* and *ser*. In other words, there would be no difference between these two verbs. In linguistic terms, this will be the place of true syntactic variation. If the category value is to the right of that point, that category is certain to be found with *estar*. If the category value is to the left of that point, that category is unlikely to be found with *estar*. Let us remember that the probability curve is steepest (i.e., almost vertical line) when there are equal probabilities. If the curve is almost vertical, we can say the predictor is certain to be found with *estar*. If the curve is not almost vertical, we can say that the predictor is leaning towards certainty.

Figure 5.10. Multilevel logistic regression of the predictor *predicate reading*. 
Table 5.14 displays a difference in the probability scale of +560% for this predictor which tells us that one of the categories does not help in predicting the appearance of *estar* but the other is almost certain to be found with *estar* for predicate reading. Figure 5.10 shows this effect visually. We can see here that when predicate reading is an individual level reading (1), it is unlikely that it will be found with *estar*, but when *predicate reading* is a stage-level reading (2), it is predicted to be found with *estar*.

From Table 5.14 above, we can see that *susceptibility to change* is not a significant predictor of *estar* and a 30% difference in the probability scale for this predictor.

![Figure 5.11. Multilevel logistic regression of the predictor susceptibility to change](image)

Figure 5.11 shows the effect of this variable and explains why it is not a significant predictor of *estar*. We can see that when the predicate is *susceptible to change* (1) and *not susceptible to change* (0); it is unlikely that we would find it with *estar*. In other words, the features *susceptible to change* and *not susceptible to change* do not help us in predicting the appearance of *estar*. 
The next predictor in the model is *experience with the referent*. From table 5.14, we can see that this variable is a significant predictor of *estar* and that the difference on the probability scale for this variable is approximately +50%.

![Figure 5.12. Multilevel logistic regression of the predictor experience with the referent](image)

From figure 5.12, we can infer that there is a difference among the categories of this variable. When the *experience with the referent* is ongoing (1), it is unlikely that we would find it with *estar*. However, when the *experience with the referent* is immediate (2) or indirect (3), it is almost certain that we would find it with *estar*.

Another significant predictor of *estar* is the variable group *adverb* (see Table 5.14). The difference in the probability scale on Table 5.14 is -15% which tells us that the presence of the adverb negatively impacts the probability of finding *estar* with an adverb.
Figure 5.13 shows that when the adverb is implied (2) it seems to approach certainty that we find it with *estar*. But when the adverb is explicit (1) and/or absent (0), it is certain that we would not find *estar*. However, it seems that the presence or absence of an adverb does not predict the appearance of *estar*.

The next significant predictor is *subject*. From Table 5.14 we can see that this predictor also negatively impacts the prediction of *estar*. From Table 5.14 we see that the difference in the probability scale is 36%.
Figure 5.14 depicts the negative impact of this predictor. When the subject is a
first order entity (1), it is unlikely that we find it withestar; and when the subject is
second order entity (2), it is certain that we would find it withestar. Because of the
negative impact on the probability scale we can say that the presence of first order
entities does not help the prediction ofestar. The presence of second order entities, on
the other hand, provides a better context for the presence ofestar.

The next significant predictor of estar is the variable group resultant state. Table
5.14 shows that this variable positively impacts the probability ofestar=1 that the
difference in the probability scale is about 58%.

Figure 5.15 shows the magnitude of this difference. In it, we can observe that
neither feature predicts the presence ofestar (i.e., the probability is zero). A 58%
difference in the probability scale tells us that, perhaps, for this targeted effect the number
of tokens is too small or that there is a confounding predictor at work, I analyze and
discuss this finding further in Chapter 6.
The next significant predictor of *estar* is the variable group *adjective class*. Table 5.14 shows that this group of variables impact positively the appearance of *estar*. The difference in the probability scale is very little (8%) which might be due to the large number of categories.

![Multilevel logistic regression of the predictor adjective class](image)

Figure 5.16. Multilevel logistic regression of the predictor *adjective class*

Figure 5.16 shows that there is a small difference between predictors and that, on one hand, adjectives of *age* (1) and *size* (2) are unlikely to be found with *estar*. On the other hand, adjectives of description/evaluation (4), description of a person(ality) (5), mental/physical state (7), sensory characteristic (8) and status (9) are certain to be found with *estar*.

The last predictor of the model is not a significant one. This variable, *gradiency*, impacts the prediction of *estar* negatively and the difference is about 5% on the probability scale (see Table 5.14).
Figure 5.17 depicts this negative impact. We can see that both gradient (1) and absolute (0) adjectives have a probability of \( \text{estar} = 1 \) of practically zero. In other words, neither type of adjective is a statistically significant predictor of \( \text{estar} \).

**Social Predictors**

Level 2 coefficients are presented as group standard deviations in the logit scale. They are interpreted following the dividing-by-4 rule presented in Chapter 4. The results yield the percent of variation on the probability scale over and above the differences explained by linguistic factors. In other words, the results represent a further explanation of variance found in the data and help us understand the influence of social factors in the prediction of \( \text{estar} \).
Table 5.15
Interpretation of level 2 parameters of predicate reading Model 36

<table>
<thead>
<tr>
<th>Level</th>
<th>Parameter</th>
<th>Variable</th>
<th>SD</th>
<th>Difference in Probability scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 02</td>
<td>(Intercept)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>1.22</td>
<td>0.305</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>0.32</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>0.30</td>
<td>0.075</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bilingualism</td>
<td>0.33</td>
<td>0.0825</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contact</td>
<td>0.00</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attitude</td>
<td>0.00</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Residual SD</td>
<td></td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.15 presents the results of related to social predictors of the multilevel logistic regression of predicate reading Model 36. As we can see, factors contact and attitude possess no standard deviation which means that these two factors do not add anything to the explanation of variance in this sample. In other words, contact and attitude do not help in explaining the phenomenon.

Age shows a standard deviation of 1.22 and a difference in the probability scale of .305. This .305 means speakers vary ± 30% on the probability scale.
Figure 5.18 shows that the younger the speaker the more production of *estar* is found in the sample. In speakers of ages 55 and up, production of *estar* is not similar to that of younger speakers. Considering the fact that *age* is a continuous variable, a 30% variation is substantial. Figure 5.20 is a histogram of the production of *estar* by age. A look at the confidence intervals of the intercepts by age shows a trend towards more accurate prediction of *estar* in data coming from younger speakers. I come to that conclusion because the 95% confidence intervals narrow considerably between the ages of 20 and 50. In this study, a narrow confidence interval translates as a more accurate prediction of the appearance of *estar*. However, there is still a lot of fluctuation within these ages.
Figure 5.19 shows that even though all ages explain the appearance of *estar* significantly (i.e., confidence intervals of their intercepts do not include zero), the pattern seems to be a narrowing of the gap between the lower bound and the upper bound. This narrowing indicates that there is more variance among speakers of ages 55 and up than the variance found among speakers of ages between 20 and 55. The narrowing of the confidence interval indicates a more accurate parameter estimation that can be translated into a higher use of *estar* in the linguistic contexts where the linguistic factors present above were significant. This accuracy in estimation of parameters in the model may be a sign of two things: first of the stabilization of the phenomenon and second of the influence of language contact. Both hypotheses will be explained in Chapter 6.

The next social variable that explains variance over and above the differences explained by linguistic factors is *gender*. This variable standard deviation is .32 that divided by 4 yields .08. In other words, gender groups vary by about 8%. In this case, females differ by about 8% from males in the production of *estar*. 

Figure 5.19. Confidence intervals of the intercept by age
Figure 5.20 shows that when compared females are producing more instances of *estar* than males. The production of *estar* by men is approximately 8% lower than the production by women (approximately 18% and 25% respectively) in the raw data.

*Education* as a social variable also helps explain variance over and above the one explained by linguistic variables alone. In this case the standard deviation of *education* is 0.3. This coefficient divided by 4 represents 7.5% of difference in the probability scale. This means that production of *estar* among education levels varies about 7.5% among groups in the probability scale.
Figure 5.21 shows this variance by plotting the production of *ser* and *estar* by education group. This figure shows an interesting social pattern. As the education level increases the overall production of *estar* decreases and the one of *ser* increases in all contexts. A chi-square of homogeneity ($\chi^2=8.2058$, df=5, $p=0.1453$) tells us that these groups are different with respect to the production of *estar*. This is an interesting finding because it might mean that the use of *estar* is becoming a sociolinguistic marker as a verb to be used in formal settings (such as an interview) when there are other verbs that convey the same meaning such as “parecía” (it seemed). The discussion regarding this variable’s influence on the predictability of *estar* is found in Chapter 6.

The last social variable that helps explain the phenomenon over and above the explanation by the linguistic factors alone is *bilingualism*. The standard deviation of this factor in the logit scale is .33 or 8.25% of difference in the probability scale. This means that bilinguals and multilinguals differ 8.25% from monolinguals.
Figure 5.22 shows that monolinguals produce more instances of estar than bilinguals and multilinguals. On the other hand, they produce fewer instances of ser than bilinguals and multilinguals. Even when bilinguals and multilinguals are combined (called bilingual for the graph), monolinguals still produce more instances of estar (see Figure 5.23).

Figure 5.23. Production of ser and estar by bilingualism
Chapter 6 follows with the discussion of how these findings fit with the general picture of sociolinguistics and the study of copula choice in Spanish.
Chapter 6
Discussion and Conclusion

Introduction

Three questions were asked at the onset of this study. The first one was what linguistic factors predict the use of *ser* and *estar + adjective* in the Spanish spoken by Costa Ricans in Limón. The second was what social factors predict the use of *ser* and *estar + adjective* in the Spanish spoken by Costa Ricans in Limón. The last question was whether the pattern of variation in the use of *ser* and *estar* could be considered a change in progress or a stable change.

This chapter is divided into three sections. The first section is the discussion of the findings to answer the research questions, the second section is a summary of the methodological contributions to the field of sociolinguistics made by this study, and the third section is the concluding remarks and recommendations for future research.

Discussion of the findings

It is worth mentioning here that one of the objectives of this study was finding a model for the prediction of *estar* in the Spanish of Limón, Costa Rica. In order to achieve this goal, I collected oral data from fifty eight randomly selected speakers of Spanish from Limón. The sample was carefully planned for the study to have a power level of 0.8. Cohen (1969, 1988, 1992a, 1992b) calls it the probability that a statistical test will yield statistically significant results. In other words, my study has eighty percent certainty that its results are statistically significant assuming necessary assumptions satisfied. This level of power allows me to generalize my findings to the population under study. To build the model I took into account theoretical accounts and the findings of previous studies on
copula choice. All viable predictors were used for the determination of sample size. Their distributions and effect sizes were determined by the findings of Aguilar-Sanchez (2007).

In order to answer these three questions, I conducted a search of all the possible factors set forth by theoretical accounts that could help predict the appearance of *estar*. These factors came from different linguistic approaches that try to explain the differences between these two verbs. Some of them came from functional approaches to Spanish in monolingual Spanish (Clements, 1988; Clements, 2006) and some came from generative approaches to the study of first language acquisition of Spanish syntax (Schmitt & Miller, 2007). However, most factors came from empirical studies on copula choice in the field of sociolinguistics in monolingual and contact Spanishes (Díaz-Campos & Geeslin, in press, 2005; Geeslin & Díaz-Campos, 2005; Geeslin & Guijarro-Fuentes, 2008; Geeslin, 2007; Guijarro-Fuentes & Geeslin, 2006) in monolingual varieties of Spanish.

Even though these factors arose from different approaches to the study of copula choice in Spanish and from different settings, the inclusion of each in my study had a methodological foundation. I chose the ones that have been proven to predict the use of *estar* and I also chose the ones that made sense from a theoretical point of view. The exercise turned out to be a fruitful one for the construction of a model to explain copula choice in the Spanish of Limón.

Key methodological issues were addressed during the design stages of this study. These issues arose while conducting the review of the previous literature. As I conducted the review, I started noticing the omission in the reporting of key procedural elements that allow a study to be generalizable to the population under study. Some of these issues also directly hindered the power of the study. These issues included the
polychotomization of continuous variables such as age, the choosing of sample sizes that were not derived from power analysis, and the inclusion of grouping factors as individual factors in their analyses.

Addressing each and every one of these issues turned out to be of paramount help in the design of this project and the calculation of the proper sample size to achieve generalizable results. Not only are these findings generalizable, they can also be compared to studies of similar power and level of generalizability in a meta-analysis of variation of copula choice in Spanish.

The discussion presented in this chapter contributes in two general areas of knowledge. The first one is directly related to the questions of the present study. This study makes two major contributions. First, it contributes to the general understanding of syntactic variation related to copula choice in Spanish and studies a variety of Spanish that has not been studied yet: Limonese Spanish. The second area is that of general research methodology in sociolinguistics. This study represents an important contribution to the methodology that has guided our research practices in search of an explanation of the role of social factors in linguistic change. The main contribution is that it provides the field with a methodology to calculate sample sizes for future research and introduces a modification to the variable rule analysis that allows studying the role of social factors without violating the mathematical assumptions made by the statistical analysis used.

In order to predict the appearance of estar in the construction estar + adjective in the Spanish of Limón, I selected, among several possible models, the model that best explained variance in the sample. An exploratory procedure of this kind allows the researcher to get very well acquainted with his/her data and allows him/her to make
accurate inferences regarding his/her construct. Thus, the statistical model I found to be
the best statistically predicting model of *estar* in the context of *copula + adjective* is:

$$Pr(estar_i = 1) = \logit^{-1}(\beta^0 + \beta^{predtype} \cdot predtype + \beta^{suschang} \cdot suschang + \beta^{expref} \cdot expref + \beta^{adverb} \cdot adverb + \beta^{subject} \cdot subject + \beta^{resultate} \cdot resultate + \beta^{adjclass} \cdot adjclass + \beta^{gradien} \cdot gradien + \alpha_{age} + \alpha_{education} + \alpha_{bilingual} + \alpha_{contact} + \alpha_{gender})$$

In linguistic terms, the variables included in this model were *predicate type*,
*susceptibility to change*, *experience with the referent*, *adverb*, *subject*, *resultant state*,
*adjective class* and *gradiency*. Social predictors included *age*, *education*, *bilingualism*,
*contact with English* and *gender*. This model helped me to answer all three questions of
this dissertation. All answers will be addressed in the forthcoming sections.

**Answer to research question one**

**The statistically significant factors**

After fitting the model, five predictors were statistically significant. These
predictors were *experience with the referent*, *adverb*, *subject*, *resultant state* and *adjective class*. Three of these predictors, *experience with the referent*, *resultant state* and *adjective class*, have proven to be very strong in the prediction of *estar* in different social and
geographical contexts (Díaz-Campos & Geeslin, 2004; Geeslin & Díaz-Campos, 2005a;
2005b; Geeslin & Guijarro-Fuentes, 2008).

*Experience with the referent* is not easily available in data collected with a
Labovian interview, as the data demonstrated, because most speakers are narrating from a
egocentric point of view which is an ongoing experience (i.e., they are narrating
experiences that are directly related to them). However, in this study and with the help of the appropriate statistical analyses that can handle smaller distributions, we saw how this variable is still and continues to be a strong predictor of *estar* in data gathered in oral interviews.

The second statistically significant predictor, *adverb*, helps us to predict when *estar* will not be used. These findings are important, because it has been argued, in the generative approach to grammar, that the use of *ser* and *estar* might be related to aspectual differences as mentioned by Schmitt and Miller (2007). See examples 6.a and 6.b.

6.a. “uno ESTÁ acostumbrado, es a ver carne en el plato” F28D07IVA
    “one is accustomed is to see meat on the plate”
    ‘one is accustomed to seeing meat on the plate”

6.b. “una señora que actualmente ES muy conocida, muy famosa” M31000IVA
    ‘a lady that these days is very kown, very famous’
    ‘a lady that is very well known and famous these days’

This finding is of importance because even though the presence or absence of a temporal adverb or temporal expression has been used as a variable (Salazar, 2007) or to test hypothesis (de Jonge, 1993) regarding the appearance of *estar*, it is the first time that this predictor appears as a statistically significant predictor of *ser* and not of *estar*. Even though Salazar (2007) found that the presence of a time adverbial favors the use of *estar* similarly to De Jonge (1993), it is not clear whether she coded for all time adverbial expressions or just a few time adverbs such as *ahora*, *ya* and *cuando*. The findings presented here validate Schmitt and Miller’s (2007) approach and represent the first study to analyze the presence or absence of any time reference within the discourse of the speaker.
The third statistically significant predictor of \textit{estar, subject}, is one of the predictors that has shown more inconsistency in the studies of copula choice in Spanish as pointed out by Geeslin and Guijarro-Fuentes (2008). Although in their study, referents, not subjects, were classified as animate or not animate, I include a broader classification that includes referents as part of the subject of the sentence. I used Clements’ (2006) classification of subjects as \textit{first-order entities} and \textit{second-order entities}. The results corroborate Clements’ (2006) predictions that \textit{first-order-entities} are found more with \textit{ser} and that \textit{second-order-entities} more with \textit{estar}.

The fourth statistically significant predictor, \textit{resultant state}, showed an unexpected behavior. In this study, it was a significant predictor of \textit{estar} but neither \textit{resultant} nor \textit{nonresultant} states seem to be found more often with \textit{estar}. The explanation of this behavior may be because there is an interaction with another predictor. This interaction may come from two different sources. The first source is linguistic and the second is social. If the first is true, the statistical analysis would have yielded a large standard error. The second source may be that there are predictors at level 2 (i.e., the social predictors) that may be influencing this predictor at level 1. Let us look at some examples:

6.a. “ahora (grupos de pescados) SON contadas las que hay,”M28C06IVA
‘now (groups of fish) are counted the ones there are’
‘now the groups of fish that you find are few’

6.a. was produced by a male bilingual speaker with a secondary incomplete level of education. The lack of gender agreement between a noun and its modifier is a common phenomenon in the bilingual speakers when speaking Spanish (see Hertzfeld, 2004 for further information regarding this and other phenomena regarding Limonese English).
6.b. “de Liverpool hasta Blanco, más o menos ESTÁ pavimentada la calle. (ahora)”

‘from Liverpool until Blanco, more or less is paved the street (now)’
‘The road is more or less paved from Liverpool to Blanco’

6.b. was produced by a female monolingual speaker with a secondary complete level of education.

6.c. “algunas familias ESTÁN mas o menos ahí estabilizadas, (ahora)”

“some families are more or less stabilized (now)”
“Some families are more or less stable (now)”

6.c. was produced by a female bilingual speaker with university complete level of education.

6.d. “porque (traer cosas) ES cansado”

“because (bring things) is tired”
“because bringing things is tiring”

6.d. was produced by a male monolingual speaker with a university incomplete level of education.

From examples 6.a to 6.d, we can see that the production of estar with resultant states may be influenced by social factors such as education, gender, and bilingualism.

I took a closer look at the distribution of estar by these social predictors to find evidence of this influence.
Figure 6.1 shows that males, both bilingual and monolingual, have a tendency to use less *estar* with resultant states whereas females, both bilinguals and monolinguals, have a tendency to use more *estar* with resultant states. The evidence presented here shows why despite the predictor *resultant state* being statistically significant, neither resultant states nor non resultant states seem to appear more frequently with *estar*. In other words, it explains the 58% difference in the probability scale noted in Chapter 5. It also serves as evidence that social factors do influence linguistic factors and that they need to be analyzed separately, a process that is facilitated by a multilevel analysis.

The last statistically significant predictor, *adjective class*, still keeps its predictive power despite the inclusion of other predictors in the model. This variable has been found to predict the appearance of *estar* in different varieties of Spanish. However some classes, which were exclusively used with ser, are now becoming predictors of *estar*. In this sample, all but two classes (*age* and *size*) are predictors of *estar*. 

Figure 6.1. Distribution of predictor resultant state by gender and bilingualism
The statistically non-significant factors

It is worth mentioning here that factors that were not statistically significant are also important in statistical modeling for prediction because they contribute to the explanation of variance found in the data. Their contribution is that they help in producing more accurate results by eliminating sources of error in the data. In other words, they help us to determine which predictors are significant.

In this study, three predictors were not statistically significant. They were predicate reading, susceptibility to change, and gradience. The predictor predicate reading is important because as pointed out in Chapter 5, it shows us that the use of ser with individual level readings of the predicate is categorical and that there is more variation when the predicate has a stage-level reading. This finding is important for this study because it represents an empirical account of a context where true variation exists between the copulas. In previous studies, the predictor predicate type has been demonstrated to be a significant predictor of estar (Díaz-Campos & Geeslin, in press; 2005; Geeslin, 2003; Geeslin & Guijarro-Fuentes, 2008; Geeslin, 2007; Guijarro-Fuentes & Geeslin, 2006), however, the new predictor proposed in this study, predicate reading, is not statistically significant for the prediction of estar. This reason of the difference may lie on the fact that this new predictor is the combination of two previously used ones: predicate type and frame of reference. In previous studies, only predicate type has shown to be significant. Frame of reference rarely was found to be significant. As I explained in Chapter 4, this might be a sign that both predictors were highly correlated and by combining them, I have eliminated the possibility of error due to lack of orthogonality and the predictor became a more accurate predictor of estar by itself, but not with other predictors as shown in Chapter 5.
The reason why the second predictor, *susceptibility to change*, was not significant may lie in the fact that for the present study, *susceptibility to change* was coded as the possibility that the characteristic of the referent might change as defined by Díaz-Campos and Geeslin (in press); therefore, any characteristic of the referent that the speaker, in the discourse, pointed out as having changed from the past to the present was coded as *susceptible to change*. The coding was done regardless of the possibility that the speaker might have seen such a characteristic as permanent or temporal. Such distinction is hard to pinpoint in events that have changed already. The use of the copula *ser* does not guarantee the impossibility of change of the characteristic.

Although by itself and in combination with other variables the predictor *gradiency* was statistically significant, it was not when included in the more complete model. This variable should not be tossed out of future studies as it is beneficial from a statistical point of view to the explanation of the variance found in the sample. Including this predictor allows the statistical model to make more accurate predictions by eliminating error; but linguistically, it seems that neither *absolute* nor *gradient* adjectives help predict *estar*.

In sum, the present study found that the linguistic factors that help predict the appearance of *estar* in the Spanish of Limón are *experience with the referent, the absence of an adverbial expression* (i.e., *adverb*), *the nature of the subject*, *resultant state*, and *adjective class*.

**Answer to Question 2**

In order to answer question number two, I switch from talking about statistical significance to talk about variation and its role in the explanation of variance in the prediction of *estar* because Level 2 predictors provide only information regarding how
the groups vary (i.e., standard deviation) , the findings help us to see patterns of a social nature regarding a syntactic phenomenon that could not be seen utilizing other statistical analyses (i.e., significance testing). See my explanation of rationale in Chapter 4, page 141.

**Predictors that did not show standard deviation**

Six predictors were entered in the model to test for their influence in the prediction of *estar*. These predictors were of a social nature. Two predictors were constructs, three demographics, and one was a survey answer. From these six variables, two did not show standard deviation among groups. This means that all groups behave in the same way which in turn means that these two variables did not influence the prediction done by the linguistic predictors. These variables were *contact* and *attitude*. The type of contact that the person had did not help the model in the prediction of *estar*. Similarly, whether the participant had a positive or negative attitude towards the preservation of indigenous languages did not help to predict *estar*.

**Predictors that showed standard deviation**

The predictors that showed standard deviation were *age*, *education*, *bilingualism*, and *gender*. These last three predictors help to answer question number two of this study. With regard to *gender*, four hypotheses are explored in search for an explanation of the results. The first three hypotheses help to reveal that in this particular geographical region two different varieties of Spanish coexist and are constraint by different social factors.

A first look at the results seems to suggest that the production of *estar* is higher in females than in males. They differ by about 8% in the probability scale. This trend seems to be normal in sociolinguistic terms because women are usually ahead in adopting change when this change does not carry a stigma (Labov, 1974).
The results seem to shed light on the gender issue because education shows a pattern previously suggested by Ortiz-López (2000). I call this the education hypothesis. In the Spanish of Limón the pattern is that the higher the level of education, the lower the use of estar and vice versa. This finding supports Ortiz-Lopez’ (2000) claim that level of education is related to copula choice.

![Figure 6.2. Distribution of gender by level of education.](image)

Figure 6.2 shows that at higher levels of education the sample of speakers is somewhat balanced; however, it is not between participants with 6 to 9 years of schooling (i.e., primary school complete and secondary school incomplete). We find more women in this group than men. This may be evidence that lack of access to formal education allows for more overall production of estar in all contexts and it seems to accelerate change.
Figure 6.3 shows that higher levels of education reported higher numbers of bilinguals whereas the lowest levels of education reported the lowest numbers of bilinguals. This piece of evidence becomes relevant for the discussion presented below.

A second possible hypothesis aside from the education hypothesis (Ortiz-Lopez, 2000) that may be formulated with a first look at the results is through a purely sociolinguistic point of view. I call it the stigma hypothesis. The fact that education is triggering the use of one copula more than the other, in this case ser, seems to be a case where stigma is being attached to the other, in this case estar. If this is the case, the explanation of the variation found in gender seems to be in one direction. Attaching stigma to the copula estar might be a recent phenomenon and women are beginning to reduce their use of estar while men are not reducing their use of it despite their level of education.
Figure 6.4 shows what seems to be support for the sociolinguistic hypothesis because women are more conservative with the use of *estar* as their education level moves up. On the other hand, males seem to increase the use of *estar* as their education level increases which means that behavior is not altered by formal education in Spanish.

A third hypothesis regarding the behavior of social predictors can also be formulated after a more careful analysis of the data. Because part of the participants are either bilingual in Spanish and another language or multilingual in Spanish and two or more other languages, the context in which this Spanish is found is one where languages are in contact. The predictor *bilingualism* showed variation in this sample. The pattern that can be observed seems to explain the phenomenon from a different perspective: the contact perspective. Bilinguals or speakers that speak more than one language use *ser* more than monolinguals; and monolinguals use *estar* more than bilinguals. This pattern
allows me to set a different hypothesis. This one is the *contact hypothesis*. Let us remember that bilingual speakers in Limón have limited access to formal education in English, but have full access to formal education in Spanish. Limonese English has one copula “be”. Even though they omit it in certain contexts, the acrolect of this language does not (Herzfel, 2004; Winkler, in progress). Therefore, another hypothesis that can be formulated is that bilinguals might be using the copula that most resembles the English “be” which is “ser” (is = es) and that gender and education level influence the impact of this predictor in the use of *estar*.

![Figure 6.5](image)

Figure 6.5. Distribution of *gender* by *bilingualism*.

Figure 6.5 shows that there are more monolingual women than men in the sample, but their numbers are somewhat balanced among bilinguals and multilinguals.
Figure 6.6 shows that monolinguals’ use of *estar* decreases as their education level increases while bilinguals’ use of *estar* increases.

All three hypotheses (*education, stigma, and contact*) seem to find support in the data. The fact that all three seem plausible means there is more to be found regarding these three nonlinguistic predictors in this variety of Spanish. A fourth hypothesis may be formulated. This hypothesis is a mixture of the *education* hypothesis set forth by Ortiz-Lopez (2000) and the *contact* hypothesis. It might be that although two varieties of Spanish are found in the same geographic region, their patterns with respect to the production of *estar* are different; therefore, producing interactions that obscure the influence of social factors in each one. I call this the *coexistence* hypothesis.

Figure 6.7 and 6.8 shows the evidence to support this hypothesis. The use of *estar* decreases as education level increases for most monolinguals speakers of Limonese Spanish.
Figure 6.7. Distribution of production of *estar* by gender and *level of education* among monolinguals.

![Monolingual Males](image1)
![Monolingual Females](image2)

Figure 6.8. Distribution of production of *estar* by gender and *level of education* among bilinguals.

![Bilingual Males](image3)
![Bilingual Females](image4)
In much different fashion, bilinguals’ use of *estar* increases as education level increases. Figure 6.8 shows evidence to support the _education_ hypothesis while Figure 6.6 shows evidence to support the contact hypothesis. The evidence for the _contact_ hypothesis is increased by the fact that the higher the education level, the higher the number of bilinguals found in the sample (see Figure 6.2 above).

![Figure 6.9](image)

*Figure 6.9. Production of *estar* by gender, bilingualism and level of education.*

Figure 6.9 shows the interaction of these variables. In other words, it explains how contact with English prompts higher use of _estar_ in the linguistic environment _copula_ + _adjective_. This also supports the claim that it is the contact with English that accelerates change (Gutiérrez, 1992; 2003; Silva-Corvalán, 1986; 1994) in bilingual varieties of Spanish and not the contact with other languages (Ortíz-López, 2000), and that access to formal education in Spanish decelerate it (Ortíz-López, 2000). It is also evidence that two different varieties of Spanish coexist in the same geographical area and that both of these varieties are constrained by different linguistic factors.
In sum, social factors such as age, gender, education, and bilinguism help to explain the phenomenon beyond what linguistic factors can explain. In this case, they have helped me determine that two varieties of Spanish lived in the same geographic area and that they are constraint by different linguistic factors.

**Answer to question 3**

In order to answer whether the linguistic phenomenon is a change in progress or a stable change, in the field of sociolinguistics, Labov (1972) and Bayley (2002) suggested the construct of apparent time. In this construct, linguistic changes are seen as differences between generations of speakers. In the general tradition of sociolinguistics, age has been constructed as age groups. However, this practice as discussed in Chapter 3 hinders the power of the study and may produce spurious results (Maxwell & Delaney, 1993). In the present study, age was analyzed as a continuous predictor at level 2 because it is demographic in nature and it groups tokens of the same speaker.

Results show that there is 30% variation in the probability scale between ages and that most of this variation is found between the ages of 25 and 55. This means that there is a pattern of differences in the production of *estar*. These differences might find an answer in the results yielded by the analyses of the other social variables and the fact that monolinguals and bilinguals show an interaction in their use of *estar* which is triggered by level of education and bilingualism.

After looking at the data by age divided between monolinguals and bilinguals, no difference was found. The same patterns, as those presented in Chapter, 5 were found for both groups of speakers. Speakers of ages below 25 seem to use *estar* to a lesser degree than those of ages between 25 and 55. Speakers of ages 55 and up use less *estar* than the other two groups. The analyses of this pattern performed in Chapter 5 allowed me to
hypothesize based on the behavior of the confidence intervals of the standard
deviation\textsuperscript{10} that the model becomes more accurate at predicting the use of \textit{estar} between
the ages of 25 and 55 and it has less accuracy in ages below 25 and above 55.

To answer question three and as suggested in Chapter 5, this might be a sign of
stabilization of the phenomenon. As is stated in the apparent time construct (Labov,
1972; Bailey, 2002), there should be a difference between older speakers and younger
speakers to determine whether a linguistic change has happened. At this moment of
analysis, \textit{estar} seems to be used with less frequency than \textit{ser} by older generations while
younger generations seem to use it more. These results allow me to hypothesize that the
change is on its way to stabilization, but that it has still has not entirely stabilized. The
stabilization of the phenomenon can be observed in that the frequencies of use in
monolingual Limonese Spanish seem to be similar to those found in other monolingual
varieties of Spanish while the ones in bilingual Limonese Spanish seem to be similar to
those found in varieties with high contact with English where \textit{ser} is the predominant
copula used.

\textsuperscript{10} Standard deviation of the intercepts at Level 2
In figures 6.10 and 6.11 we find further evidence to the hypothesis of stabilization of the phenomenon in Spanish. When we separate the speakers by age and bilingualism, the patterns show differences in the ages. Monolingual Limonese Spanish seems not to
differ in the distribution of production of *estar* as a group. Bilinguals on the other hand, show a pattern of increase in use of *estar* in speakers less than 55 years of age and speakers under the age of 25 behave like monolinguals under the age of 25. This pattern could be caused by the fact that the younger generations tend to be fully bilingual and have more access to formal education in both languages. The older generations were dominant English speakers and had less access to formal education in Spanish. The generation between 25 and 55 are more varied in terms of bilingualism and their access to education was not consistent. While some had formal education in English and Spanish and spoke both languages at home, others only had Spanish formal education and spoke English (Herzfeld, 2004).

**Methodological contributions**

The present work has made two major methodological contributions to the field of sociolinguistics. First, it presented a methodology to determine sample sizes based on power for the study of sociolinguistic data. By calculating the proper sample size for power, we will be able to generalize our findings to the population at large and it will make our studies comparable across varieties of Spanish. Second, it provides the field of sociolinguistics with a modification to the variable rule analysis that allows us to see how social predictors help explain the variance, from a statistical point of view. By being able to do this, we will be able to draw more accurate inferences of the social behavior of language. This data is hierarchical in nature and studying it as aggregated data may lead us to wrong conclusions. Differences due to social factors are obscured when the data is aggregated; therefore, a multilevel analysis in the context of logistic regression is a more powerful method for the prediction of the influence of social predictors on linguistic phenomena.
A multilevel analysis of variance at level 1 (i.e., linguistic predictors) allowed me to determine that predictors that have been found to be strong predictors of *estar* in other varieties of Spanish are strong predictors of *estar* in Limonese Spanish. These predictors are *experience with the referent*, *adverb*, *subject*, and *adjective class*. These predictors are only strong predictors of *estar* when they are in conjunction with predictors such as *predicate reading*, *susceptibility to change*, *resultant state*, and *gradiency*. Because most of the variance explained, from a statistical point of view, came from linguistic factors, a multilevel analysis allowed me to determine that variation of copula choice in Limonese Spanish is first and foremost a syntactic phenomenon constraint by discursive and pragmatic features in accord with previous studies of copula choice in Spanish.

A multilevel analysis at level 2 (i.e., social predictors) allowed me to discover that two varieties of Spanish can live in the same geographical area and be constraint by different social factors. It allowed me to see how contact with formal education, levels of bilingualism, and gender led me to this finding. Monolingual Spanish behaves differently than bilingual Spanish because of access to formal education, or lack thereof, and levels of bilingualism. A multilevel analysis at level 2, allowed me to determine that the extension of *estar* seems to be stable in the monolingual variety of Spanish while it is still ongoing in the bilingual variety of Spanish with younger speakers approximating the use of monolingual speakers.
Future research and Conclusion

Future research should focus on the understanding of copula choice in the two different varieties of Spanish (i.e., monolingual and bilingual) that are found in Limón, Costa Rica. Further research is needed to tease out the different linguistic features that help predict copula choice for each of these varieties of Spanish. The influence of language policies and how they have affected language change deserve attention from researchers. In other words, we need to look for further evidence of the influence of formal education in the processes of language change.

The present study would benefit with a more fine-tuned instrument to measure levels of contact in order to measure it more accurately. It will also benefit with a more accurate index to classify socioeconomic status of the participants. Further analysis of the effects of predictors at level 02 is needed to determine the effects (random and fixed) of each predictor with respect to the prediction of estar. These analyses were not conducted because the scope of this work did not allow it, but future studies should explore such effects.

In conclusion, a careful sample size planning procedure under the power analytic framework has proven to be of paramount importance to the study of sociolinguistic phenomena. Not only are the results of this study reliable, they are also generalizable to the population under study. In other words, the results of this study are representative of the phenomenon in the population. The modification of the analysis that we normally carried out on the data was made to include new advances in statistical tests has resulted in a more powerful way to analyze sociolinguistic phenomena, especially the study of syntactic variation. The inclusion of new and more powerful statistical tests that is able to handle the structure of the data collected for this study (i.e., hierarchical) to the analysis
of linguistic phenomena represents a step forward in our scientific journey regarding the study of language variation. Furthermore, studying copula choice in the context of copula + adjective in the Spanish of Limón has proven to be a good example where two different varieties of Spanish coexist. Limón has proven once more to be a good case scenario for the study of linguistic processes in real time which otherwise would take hundreds of years to be observed (Winkler, 1998).
Bibliography


Falk, J. (1979). *Ser y estar con atributos adjetivales: anotaciones sobre el empleo de la cópula en catalán y en castellano*. Uppsala

Stockholm: Almqvist & Wiksell International.


Inman, C., Mesa, N., Oleas, R., & de los Santos, J. J. (1997). *Impacts on Developing Countries of Changing Production and Consumption Patterns in Developed Countries: The Case of Ecotourism in Costa Rica*. San Jose, Costa Rica: INCAE.


Appendix 1
Monte Carlo Simulation for the omnibus effect

```r
#Monte Carlo simulation
#Loop to get the power calculations

J.values<- c(20, 30, 60, 100, 200, 300)
n.sims.values <- rep(1000,9)
K.values <- c(25, 30, 35)
power.values <- array (NA, c(length(J.values),length(K.values)))
for (i1 in 1:length(J.values)) {
  for (i2 in 1:length(K.values)) {
    cat ("computing power calculation for J =", J.values[i1], ", K =", K.values[i2], "\n")
    power.values[i1,i2] <- estar.power (J=J.values[i1], K=K.values[i2], n.sims=n.sims.values[i1])
    cat ("power =", power.values[i1,i2], "\n")
  }
}

#01-21-08 Started at 7:40 a.m. and Ended at 2 a.m.

#plot all the cures

plot (c(0,max(J.values)), c(0,1), xaxs="i", yaxs="i", xlab="number of speakers", ylab="power", type="n")
for (i2 in 1:length(K.values)) {
  lines (c(0,J.values), c(.025, power.values[,i2]))
}

save.image("E://Software//R-Sessions//01-21d-08")
```
Appendix 2
Monte Carlo Simulation for the targeted effect

#This function to test for power for the variable coefficient of the variable "gradiency" based on Gelman and Hill (2007). Code taken from their chapter on sample size

```r
estar.power <- function (J, K, n.sims=1000) {
  signif <- rep (NA, n.sims)
  for (s in 1:n.sims) {
    estar.data <- estar.fake (J, K)
    lme.power <- lmer(estar ~ subject + adjectiveclass + understructure + gradiency + predicatetype + succhange + frameref + expreferent + resultstate + (1 |gender1) + (1 | age1) + (1 | educ1) + (1 | biling1),
      family=binomial(link="logit"), data=estar.data)
    theta.hat <- fixef(lme.power)
    theta.se <- se.fixef(lme.power)
    signif[s] <- ifelse (sqrt(theta.hat^2) - (2*theta.se) > 0,1,0)
  }
  power <- mean (signif)
  return (power)
}

estar.power(J=60, K=30, n.sims=10)  # this code tests the function estar.power

save.image("E://Software//R-Sessions//01-21d-08")

#Monte Carlo simulation
#Loop to get the power calculations

J.values<- c(20, 30, 60, 100, 200, 300)
n.sims.values <- rep(1000,9)
K.values <- c(25, 30, 35)
power.values <- array (NA, c(length(J.values),length(K.values)))
for (i1 in 1:length(J.values)){
  for (i2 in 1:length(K.values)){
    cat ("computing power calculation for J =", J.values[i1], ", K =", K.values[i2], "\n")
    power.values[i1,i2] <- estar.power (J=J.values[i1], K=K.values[i2], n.sims=n.sims.values[i1])
    cat ("power =", power.values[i1,i2], "\n")
  }
}

#01-21-08 Started at 7:40 a.m. and Ended at 2 a.m.

#plot all the cures

plot (c(0,max(J.values)), c(0,1), xaxs="i", yaxs="i", xlab="number of speakers", ylab="power", type="n")
for (i2 in 1:length(K.values)){
  lines (c(0,J.values), c(.025, power.values[i2,])
}
```

#01-21-08 Started at 7:40 a.m. and Ended at 2 a.m.
Appendix 3
Sample Transcription

Fecha de inicio: MZO 09 DE 2009, 10:00 a.m.
Transcriptor: 02
Entrevistado: F38A12-GIN
Fecha de la entrevista:

START

E= Entrevistador
S= Sujeto o entrevistado
O= otra persona

// = silencio o pausa
[XXX] = palabras que no puedes entender
[...] = silencio prolongado
“beach ”= palabras en otro idioma
[Risa] = risa
[Ruido] = ruido que no permite la transcripción
* = palabra transcrita como se escucha/no se entiende el significado. “Cho” “Diay”

Transcribir como se oye, no corregir, acentuación pero puedes agregarla.

START 09032009 10:00 a.m.

E: OK, este es el micrófono entonces se lo voy a poner aquí cerquita para que me pueda concentrar la voz y salga bien la grabación, Podrías describirme con detalle su niñez? Dónde jugaba? A dónde iba a la escuela? Los lugares, describirme todo eso. Todo lo que hacia desde niña, todo lo que se acuerde.

S: Bueno, yo niña iba a la escuela de varones, en el centro y me gustaba jugar así entre las chiquillas bate, quedo y de vez en cuando nos escapábamos, nos íbamos a las playas, nos íbamos al cine y Ahi andábamos por todo lado jugando, jugando bolinchas, chiquitillos, nos íbamos a, jugábamos escondido con los compañeros y con los chiquitos del barrio
E: Mjmm.
S: Y yo si pase, ves, cositas así.
E: Vivía aquí o vivían en...
S: Vivíamos en Cerro Mocho.
E: Cerro mocho?
S: Mjmmm.
E: Y me podría describir Cerro Mocho? Que no he estado ahí yo.

S: Cerro Mocho es un barrio bien lindo, por ahí por el colegio de, del centro aquí por el [xxx] por ahí.
E: Ajá.
S: Ajá. Ahi son casas bonitas. Ah! bueno, y también mi mamá nos mandaba a nosotros a vender empanadas.
E: Ajá.
S: A los colegios, nos íbamos todas las noches a las cinco, y vendíamos las empanadas, y veníamos veníamos corriendo por mas empanadas, contentas, “mama, se vendieron las empanadas, denos otras” y nosotros nos íbamos con mas empanaditas y se venía to’as las empanadas
E: umjum
S: cerca del colegio, ahí por la guardería de Limon, infantil, ahí, muchos vecinos todos eramos muy amigables, vecinos, amigos y hermanos, todos nos compartíamos todo.
E: Mjmm.
JORGE AGUILAR-SÁNCHEZ  
3450 N. Jayne Ann Ct.  
Bloomington, IN 47404  
Phone (812) 361-0173 (m); (812) 876-0521 (h)  
e-mail: aguilar.jorg@uwlax.edu

EDUCATION

Indiana University, Bloomington, IN  
Dissertation Title: Syntactic Variation: The case of copula choice in Limón, Costa Rica  
Dissertation Committee: Manuel Díaz-Campos (Chair), Kimberly Geeslin, Ken Kelley, Beverly Hartford, Harry L. Gradman.  

2005 – 2006 M.A. in Hispanic Linguistics  
Indiana University, Bloomington, IN  
Specific areas: Sociolinguistics, Second Language Acquisition, and Syntax  
MA exams committee: Manuel Díaz-Campos (Chair), Kimberly Geeslin, Joseph Clancy Clements

Indiana University, Bloomington, IN  
Specific Area: ESL Grammar  
Thesis: A revised approach to teaching the English tense and aspect system based on empirical evidence from the field of second language acquisition and a critical review of current pedagogical practice.  
Coursework finished in 2002, Thesis defended in 2004  
MA Thesis committee: Harry L. Gradman (Chair), Beverly Hartford, Daniel J. Reed.

1997 – 2000 Licenciatura (MA equivalent) en Lingüística Aplicada (finished coursework)  
Universidad Nacional, Heredia, Costa Rica

1991 – 1995 Bachiller en la Enseñanza del Inglés  
Universidad Nacional, Heredia, Costa Rica

CURRENT POSITION

2009 – Present  Assistant Professor of Spanish, Department of Modern Languages, University of Wisconsin-La Crosse, WI

PREVIOUS POSITIONS HELD

2004 – 2009 Associate Instructor of Spanish, Department of Spanish and Portuguese, Indiana University

1998 – 2009 Associate Professor of Applied Linguistics, Escuela de Literatura y Ciencias del Lenguaje, Universidad Nacional, Costa Rica

2003 – 2004 Research Assistant, Language Testing Laboratory, Department of TESOL and Applied Linguistics, Indiana University

2001 – 2004 Spanish Foreign Language Instructor. People’s University, City of Bloomington Parks and Recreation Department, Bloomington-IN


RESEARCH INTERESTS

EXPERIMENTAL AND VARIATIONIST SOCIOLINGUISTICS: Syntactic variation, sample size for variationist analyses, the operationalization and quantification of language attitudes an/or qualitative data, corpus contemporary linguistics.
LANGUAGE CONTACT AND BILINGUALISM: Language contact and language change, development of new varieties of Spanish and English, language standardization and language policies.

SECOND LANGUAGE ACQUISITION: The acquisition and teaching of tense, aspect, and mood systems; the acquisition of syntactic/pragmatic structures, the acquisition of language variation.

RESEARCH DESIGN: Sample size planning for linguistic analyses, quantification methods of qualitative data, computer assisted data collection methods.

SECOND LANGUAGE TEACHING METHODOLOGIES: Processing Instruction and input processing, communicative language teaching, technology in the foreign/second language classroom, teacher training both pre- and in-service.

PUBLICATIONS (*= refereed, # = Conference Proceedings, & = Book Chapters, **= Thesis)


REFEREED CONFERENCE PRESENTATIONS

2009 How many speakers? How many tokens?: A methodological contribution to the study of variation. New Ways of Analyzing Variation 38 (NWAV 38), University of Ottawa, Ottawa, Canada.


2007 A new approach to the study of the acquisition of pragmatic competence: The case of Spanish tú and usted. Georgetown University Round Table on Languages and Linguistics (GURT), Washington, DC.

2007 The use of Spanish Ser and Estar + adjectives: A sociolinguistic pilot study on the oral Spanish of Costa Rica. 5th Hawaii International Conference on the Arts and Humanities, Hawaii International Conferences, HI.


2005 Costa Rican Englishes: The struggle, induced by language attitudes, of an inner circle against an expanding circle variety. International Association for World Englishes Conference 2005, Purdue University, West Lafayette, IN.
2005 (with Erin McNulty and James F. Lee) *Is Computer Assisted Language Learning Superior to Traditional in-class Language Learning*. Department of Spanish and Portuguese Graduate Student Advisory Committee Conference, Indiana University, Bloomington, IN.

2004 *Teaching the English tense-aspect system: An overview of Empirical Evidence in SLA, a textbook analysis, and a pedagogical multi-level approach for the creation of form-meaning associations*. INTESOL Annual Conference, Carmel, IN.


2002 *Teaching English Supra-Segmentals through POPular Songs*. INTESOL Annual Conference, Indianapolis, IN.

2001 *English in Costa Rica*. Indiana First World Englishes Round Table. Indiana University, Bloomington, IN.

**INVITED TALKS/WORKSHOPS**


**RESEARCH/FIELD WORK EXPERIENCE**

- **Summer 2008** Research Assistant
  Dr. Elizabeth Winkler
  Collected oral data in Limón, Costa Rica

- **2007 – 2008** Field Research
  Dissertation Field Work
  Collected 57 20-minute oral interviews

- **Summer 2006** S105 Intensive Beginner Spanish material creator and proofreader. Indiana University
  Dr. Juan Manuel Soto Arribi
  Created items, embedded audio for the Ancla project of the Department of Spanish and Portuguese. Proyecto Ancla is a technology-enhanced instructional project developed by the Department of Spanish and Portuguese at Indiana University, Bloomington. Its purpose is to enhance the content and evaluation aspects of language and culture courses through innovative uses of technology, customized for learners of Spanish as an additional tool for classroom instruction.

  Prof. Daniel Reed and Prof. Charles Stansfield
  Created items, field-tested the adaptation to Spanish of the Modern Language Aptitude Test for Elementary Schools (MLAT-E to MLAT-ES), applied and coordinated the validation of the MLAT-ES in Costa Rica.

- **2002-2004** Research Assistant. Language Testing Laboratory. Indiana University
  Prof. Harry L Gradman and Prof. Daniel Reed.

- **Summer 2002** Research Assistant. Speech and Hearing Laboratory. Indiana University
  Prof. Raquel Andersen
  Collected oral data from primary school children and transcribed data collection sessions using CHILDES code.

**TEACHING INTERESTS**

**ADVANCED LINGUISTICS:** Sociolinguistics, Language Variation and Change, Language Contact, Second Language Acquisition, Variationist SLA, Functional Syntax, Non-Traditional Grammars, World Englishes.

**ADVANCED AND INTRODUCTORY RESEARCH METHODOLOGY:** Statistical Analyses for Applied Linguistics; Sample Size Planning for Applied Linguistics, Data Collection Techniques and Procedures (Corpora) for Applied Linguistics, Research Methodology for Classroom-based research, Classroom-based Research.

INTRODUCTORY AND ADVANCED TEACHER TRAINING: Spanish and English Teaching Methodologies for Teachers, Spanish and English Syntax for Teachers, Spanish and English Pronunciation for Teachers.

ACADEMIC/COMMUNITY SERVICE

Committee/Board

2007  Comité de Asuntos Académicos Congreso Internacional en Lingüística Aplicada
      Selection Committee
      Escuela de Literatura y Ciencias del Lenguaje, Universidad Nacional
      Heredia, Costa Rica

2004 – 2005  Graduate Student Representative
             Indiana TESOL (INTESOL)
             Indianapolis, IN
             Graduate Student Advisory Committee
             Member, Conference and Workshop Organizer
             Department of Spanish and Portuguese, Indiana University
             Bloomington, IN

2002 – 2005  Board Member and Web Master
             La Casa Latino Cultural Center, La Central Latina, Centro Comunal Latino, and Hoosier-Latino
             GLBT HIV-prevention group
             Indiana University, Bloomington, IN

2001 – 2003  International Center Assistant Coordinator
             Activity coordinator, webpage designer. International Student Orientation Coordinator, Founder of
             the Latin American Student Association.
             Leo R. Dowling International Center, Indiana University
             Bloomington, IN

2000  Teaching Practicum. Centro de Investigación en Docencia y Educación/Escuela de Literatura y
      Ciencias del Lenguaje, Universidad Nacional, Heredia, Costa Rica
      Supervisor and advisor.
      Student: Javier Calvo
      Institution: Colegio San José de la Montaña, Heredia, Costa Rica
      Student: Daisy Wallace
      Institution: Colegio San José de la Montaña, Heredia, Costa Rica

1999 – 2000  Undergraduate Student Academic advisor. Escuela de Literatura y Ciencias del Lenguaje. Heredia,
             Costa Rica.

Review/Referee

2007  Abstract Reviewer for the 1er Congreso en Lingüística Aplicada, UNA-CR

2007  Paper Reviewer for the Proceedings of the 1er Congreso en Lingüística Aplicada, UNA-CR

2005  Abstract Reviewer for the INTESOL 2006 Conference, INTESOL

Panel Chairperson

2006  The Graduate Student Experience: Challenges and Achievements. INTESOL Conference 2005

Conference/Workshop Organization

2007  1er Congreso en Lingüística Aplicada, Universidad Nacional-CR

2005  INTESOL Conference 2005

2004  GSAC Conference in Literature and Linguistics, Indiana University
2003 – 2006  Foreign/Second Language Share Fair, Indiana University
2001 – 2004  International Student Orientation, Indiana University

HONORS AND AWARDS

2003 – present  Associate Instructorship to finish Ph.D. in Linguistics and subsequent M.A. and Ph.D. in Hispanic Linguistics. Department of Spanish and Portuguese. Indiana University
2007  Agapito Rey Travel Grant to participate in the Hawaii International Conference on the Arts and Humanities ($400). Department of Spanish and Portuguese. Indiana University

PROFESSIONAL MEMBERSHIPS

Teachers of English to Speakers of Other Languages, Inc. (TESOL)
Moder Languages Association (MLA)
Indiana Teachers of English to Speakers of Other Languages (INTESOL)

LANGUAGES

English  Near-native speaking, reading and writing
Spanish  Native speaking, reading and writing
Latin  Good reading proficiency
French  Good reading proficiency, beginner speaking
Albanian  Knowledge of structure.

RELEVANT COMPUTER SKILLS FOR RESEARCH

R  Open source, multi-platform package for statistics computing and graphics. Proficient in scripting.
SPSS  Predictive analytic software. Proficient user.
GoldVarb  Methodological tool for variationist sociolinguistics. Proficient user.
WaveSurfer  Open source tool for sound visualization and manipulation. Proficient user.
Speech Analyzer  Computer program for acoustic analysis of speech sounds. Proficient user.

RELEVANT COMPUTER SKILLS TO INCORPORATE TECHNOLOGY IN THE CLASSROOM

Art Design for Foreign Language Teaching on-line.
Web-page design for Foreign Language Teaching online (Dreamweaver, Flash, Fireworks, Publisher, among others)
ONCOURSE CL management (Sakai) and Virtual Classroom management (Blackboard and similar software).
Computer programs such as Microsoft Access, Word, Excel, Access, and Power Point.
Office Management and Filing systems.