

L3 PORTUGUESE BY SPANISH-ENGLISH BILINGUALS: COPULA
CONSTRUCTION USE AND ACQUISITION IN CORPUS DATA

by

Adriana Picoral

Copyright © Adriana Picoral 2020

A Dissertation Submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN SECOND LANGUAGE
ACQUISITION AND TEACHING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2020

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Adriana Picoral

titled: L3 Portuguese by Spanish-English bilinguals: copula construction use and acquisition in corpus data

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.



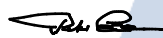
Shelley L Staples

Date: Apr 26, 2020



Ana Maria Carvalho

Date: Apr 27, 2020



Peter M Ecke

Date: Apr 26, 2020

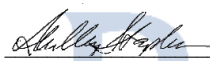


Michael Hammond

Date: Apr 27, 2020

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.



Shelley L Staples

Date: Apr 26, 2020

Department of English/Second Language Acquisition and Teaching

ARIZONA

ACKNOWLEDGEMENTS

First of all, I wish to express my appreciation to my supervisor Shelley Staples for all her tireless guidance and encouragement over the past four years. Thank you for making me a better researcher. Her infectious enthusiasm for language study and her research ethics have been a major driving force in my graduate career. I also wish to express my deepest gratitude to Ana Carvalho, who was always there to support me but also push me to think more critically about my research. Her willingness to give her time so generously has been very much appreciated.

I would like to extend my sincere gratitude to the other members of my committee, Mike Hammond and Peter Ecke. I am extremely grateful to them for their expert and valuable critiques of this research work. Thank you for helping me become a better scholar and writer.

I owe so much of the success of this project to the students and instructors of the Portuguese Language department at the University of Arizona, for enthusiastically allowing me to come into their classes and collect their assignments for my dissertation project. This project would certainly not have been possible without their support.

I also place on record my thanks to Bruna Sommer-Farias and Aleksey Novikov, who embarked on MACAWS with me from its very inception. In addition, they, along with so many other friends, made my time as a graduate student at the University of Arizona more positive and enjoyable.

Graduating during a pandemic has been a challenging process, that has been attenuated by the friendship of wonderful female colleagues. Here I thank Bruna again, for being there since before day one. To Rachel LaMance and Elif Burhan Horasanlı, for being the best conference company, listening to my rants, and validating my frustrations. To Mariela Lopez and Míriam Rodríguez Guerra, for being there to celebrate every win, and commiserate over setbacks.

Last, and certainly not least, I wish to acknowledge the support and great love of my partner Carlos. I thank him for his patience, and his willingness to listen through the ups and downs of my writing this dissertation. I love you. Thank you for reminding me that this is just a start to a wonderful new journey together.

DEDICATION

For Carlos and Daisy.

*Para minha mãe Suzana, sempre presente mesmo longe.
Para minha vó Haydée, que partiu mas mesmo assim fica.*

TABLE OF CONTENTS

LIST OF FIGURES	8
LIST OF TABLES	11
ABSTRACT	14
CHAPTER 1 INTRODUCTION	15
1.1 Research Questions	17
1.2 Overview	17
CHAPTER 2 LITERATURE REVIEW	19
2.1 Introduction	19
2.2 L3 Acquisition	19
2.2.1 L3 Acquisition Models	20
2.2.2 Other L3 Acquisition Models	28
2.2.3 Initial Transfer vs. L3 Development	30
2.2.4 Summary of L3 Acquisition Models	31
2.3 Copula Structure	32
2.3.1 Nominal Predicate	35
2.3.2 Adjectival Predicate	37
2.3.3 Prepositional Predicate	40
2.3.4 Extension of <i>estar</i>	42
2.3.5 Verbs Other Than <i>ser</i> and <i>estar</i>	43
2.4 Copula Acquisition	45
2.4.1 Copula Acquisition in Spanish	45
2.5 Conclusion	48
CHAPTER 3 METHODS	50
3.1 Introduction	50
3.2 Baseline Corpora	54
3.2.1 Portuguese	54
3.2.2 Spanish	56
3.2.3 English	59
3.3 Learner Corpus	61
3.4 Analysis	66

TABLE OF CONTENTS – *Continued*

3.4.1	Data Extraction	66
3.4.2	Quantitative Analysis	67
3.5	Conclusion	73
CHAPTER 4 RESULTS 1 – EMBEDDINGS FOR SPANISH CORPORA		75
4.1	Introduction	75
4.2	Spanish in Arizona, Mexico, and Spain	76
4.2.1	ESTAR Preference with Adjectives	77
4.3	Conclusion	85
CHAPTER 5 RESULTS 2 – EMBEDDINGS FOR BASELINE CORPORA		87
5.1	Introduction	87
5.2	Southern Arizonian Spanish vs. Brazilian Portuguese	88
5.2.1	ESTAR Preference with Adjectives	88
5.2.2	ESTAR Preference with Intensifiers (Adverb)	94
5.2.3	ESTAR Preference with EM/EN Prepositional Predicates	95
5.2.4	Summary of Differences: Arizona vs. Brazil	98
5.3	American English vs. Brazilian Portuguese	99
5.4	Conclusion	103
CHAPTER 6 RESULTS 3 – LOGISTIC REGRESSION FOR BASELINE CORPORA		105
6.1	Introduction	105
6.2	Adjectival Predicates	108
6.2.1	Adjectives of Description	108
6.2.2	Adjectives of Evaluation	112
6.2.3	Predicates with Verbal Adjectives	113
6.3	Prepositional EM Predicates	114
6.4	Conclusion	118
CHAPTER 7 RESULTS 4 – EMBEDDINGS FOR L3 PORTUGUESE CORPUS		120
7.1	Introduction	120
7.1.1	ESTAR Preference with Adjectives	121
7.1.2	ESTAR Preference with EM Prepositional Predicates	127
7.2	Conclusion	129
CHAPTER 8 RESULTS 5 – LOGISTIC REGRESSION FOR L3 PORTUGUESE CORPUS		132
8.1	Introduction	132
8.2	Adjectival Predicate	133

TABLE OF CONTENTS – *Continued*

8.2.1	Adjectives of Description	134
8.2.2	Adjectives of Evaluation	137
8.2.3	Verbal Adjectives	140
8.3	Prepositional EM Predicates	145
8.4	Conclusion	150
CHAPTER 9 DISCUSSION		154
9.1	Introduction	154
9.2	Research Question #1	155
9.2.1	Adjectival Predicates	156
9.2.2	Prepositional Predicates	159
9.2.3	Nominal Predicates	159
9.3	Research Question #2	160
9.3.1	Which language is the source of initial transfer for each Spanish-English bilingual group?	163
9.3.2	At what point (i.e., first, second or third semester) do L3 Portuguese patterns of copula use become most similar to L1 Portuguese (as opposed to most similar to L1 Spanish or L1 English)?	164
9.3.3	How similar are the L3 development paths across the three Spanish-English bilingual groups?	166
9.4	Limitations	167
9.5	Implications	168
9.6	Future Research	170
APPENDIX A Adjective Restrictions in Spanish		172
A.1	Adjectives used with <i>ser</i> only	172
A.2	Adjectives used with <i>estar</i> only	173
REFERENCES		174

LIST OF FIGURES

3.1	Percentage of undergraduate Hispanic students enrolled at University of Arizona across the years (2009-2017).	65
3.2	Density distribution of ESTAR Preference with all lexical items across languages (N=474,572). Charts are centered on zero. Positive numbers (right from zero) represent a preference for ESTAR. Negative numbers (left from zero) represent a preference for SER. Distributions are mostly symmetrical, uni-modal, and normally distributed for all corpora	72
4.1	ESTAR Preference with adjectives of age across Spanish baseline oral corpora. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	78
4.2	ESTAR Preference with adjectives of size across Spanish baseline oral corpora. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	81
4.3	ESTAR Preference with adjectives of evaluation across Spanish baseline oral corpora. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	84
5.1	ESTAR embedding preference with adjectives of age across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	89
5.2	ESTAR embedding preference with adjectives of size across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	91
5.3	ESTAR embedding preference with adjectives of physical appearance across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	93

LIST OF FIGURES – *Continued*

5.4	ESTAR embedding preference with EM prepositional predicates across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	95
5.5	ESTAR embedding preference with EM prepositions across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	97
5.6	Estimated distance of adjectives in predicate position for English copula BE across cognate with Portuguese status for the CORE corpus. Blue bars represent 95% confidence intervals. Higher numbers represent bigger distances.	100
5.7	Estimated distance of adjectives in predicate position for English copula BE across cognate with Portuguese status for the BangorTalk Miami corpus. Blue bars represent 95% confidence intervals. Higher numbers represent bigger distances.	101
5.8	Estimated distance of adjectives in predicate position for English copula BE across cognate with Portuguese status for the Cambridge corpus. Blue bars represent 95% confidence intervals. Higher numbers represent bigger distances.	102
6.1	Logistic regression results (in probability estimates) for ESTAR vs. SER with 95% confidence intervals.	109
6.2	Logistic regression ESTAR selection probability estimates.	111
6.3	Logistic regression ESTAR selection probability estimates for evaluation adjectives.	113
6.4	Logistic regression ESTAR selection probability estimates for verbal adjectives.	115
6.5	Logistic regression ESTAR selection probability estimates EM prepositional predicates.	117
7.1	ESTAR Preference with age adjectivies across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	122
7.2	ESTAR Preference with size adjectivies across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	123
7.3	ESTAR Preference with physical appearance adjectivies across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	125

LIST OF FIGURES – *Continued*

7.4	English COGNATE preference for adjectives across L1 English and Spanish Heritage groups. Values above zero represent a preference for English COGNATE, values below zero represent a preference for Spanish COGNATE. L1 Spanish group is not represented due to their exclusive preference for Spanish COGNATES.	127
7.5	ESTAR Preference with EM prepositional predicates across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.	129
8.1	Logistic regression ESTAR selection probability estimates. The dotted red line marks lower confidence for Brazilian Portuguese, for reference.	136
8.2	Logistic regression ESTAR selection probability estimates. The dotted yellow line marks lower confidence for Brazilian Portuguese and the black dotted line marks lower confidence for Spanish, for reference.	138
8.3	Logistic regression ESTAR selection probability estimates for evaluation adjectives. The dotted red line marks lower confidence for Spanish, for reference.	141
8.4	Logistic regression ESTAR selection probability estimates for verbal adjectives. The dotted red line marks lower confidence for Brazilian Portuguese, for reference.	144
8.5	Logistic regression ESTAR selection probability estimates for verbal adjectives. The dotted red line marks lower confidence for Brazilian Portuguese, for reference.	146
8.6	Logistic regression ESTAR selection probability estimates EM prepositional predicates. The dotted red lines mark lower confidence for Arizonian Spanish and upper confidence for Brazilian Portuguese, for reference.	149
8.7	Logistic regression ESTAR selection probability estimates EM prepositional predicates. The dotted black line mark lower confidence for Arizonian Spanish and the dotted yellow line marks upper confidence for Brazilian Portuguese, for reference.	151

LIST OF TABLES

3.1	Summary of corpora used in this dissertation.	54
3.2	Distribution of participants by age group and gender for C-ORAL-BRASIL (i.e., L1 monolingual Brazilian Portuguese)	55
3.3	Distribution of participants by education level and gender for C-ORAL-BRASIL (i.e., L1 monolingual Brazilian Portuguese)	55
3.4	Word count distribution across genres in the web corpus of Brazilian Portuguese.	56
3.5	Distribution of participants by age group and gender for CESA (i.e., bilingual Spanish in Southern Arizona).	57
3.6	Distribution of participants by education level and gender for CESA (i.e., bilingual Spanish in Southern Arizona).	57
3.7	Distribution of participants by age group and gender for PRESEEA subcorpora (i.e., Mexico and Spain).	58
3.8	Distribution of participants by education level and gender for both PRESEEA subcorpora (i.e., Mexico and Spain).	59
3.9	Summary of the oral Spanish baseline corpora used.	59
3.10	Distribution of participants by age group and gender for the Cambridge Spoken corpus.	60
3.11	Distribution of participants by education group and gender for the Cambridge Spoken corpus	61
3.12	L3 Portuguese learner corpus summary of document count, total word count, and mean document length in words per course.	62
3.13	Age distribution of L3 Portuguese corpus participants.	63
3.14	Year in school distribution of L3 Portuguese corpus participants.	64
3.15	Distribution of copula instances across corpora.	67
4.1	Adjectives of age in the Spanish corpora.	77
4.2	Linear regression results with 95% confidence intervals for adjectives of age across Spanish baseline oral corpora.	78
4.3	Adjectives of size in the Spanish corpora.	80
4.4	Linear regression results with 95% confidence intervals for adjectives of size across Spanish baseline oral corpora.	81
4.5	Adjectives of evaluation in the Spanish corpora.	83
4.6	Linear regression results with 95% confidence intervals for evaluation adjectives across Spanish baseline oral corpora.	84
4.7	Summary of word embedding results for the Spanish corpora.	86
5.1	Linear regression results with 95% confidence intervals for adjectives of age across Arizona and Brazil oral corpora.	89

LIST OF TABLES – *Continued*

5.2	Linear regression results with 95% confidence intervals for adjectives of size across Arizona and Brazil oral corpora.	90
5.3	Linear regression results with 95% confidence intervals for adjectives of physical appearance across Arizona and Brazil oral corpora.	92
5.4	Linear regression results with 95% confidence intervals for intensifiers across Arizona and Brazil oral corpora.	94
5.5	Linear regression results with 95% confidence intervals for EM prepositional predicates across Arizona and Brazil oral corpora.	96
5.6	Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position of copula be in the CORE corpus.	100
5.7	Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position of copula be in the BangorTalk Miami corpus.	101
5.8	Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position of copula be in the Cambridge corpus.	102
5.9	Summary of word embedding results for baseline corpora.	103
6.1	Logistic regression results with 95% confidence intervals across predicate types. First number under each L1 is the estimate mean ESTAR probability. Numbers inside the square brackets are 95% confidence intervals.	108
6.2	Logistic regression results for ESTAR vs. SER with description adjectives across L1s.	110
6.3	Logistic regression results for ESTAR vs. SER with evaluation adjectives across L1s, with ESTAR as reference.	112
6.4	Logistic regression results for ESTAR vs. SER with verbal adjectives across L1s.	114
6.5	Logistic regression results for ESTAR vs. SER with EM prepositional predicates in Portuguese and Spanish, with ESTAR as reference.	116
6.6	Summary of logistic regression results for baseline corpora. All results are significantly different from a chance probability.	118
7.1	Linear regression results with 95% confidence intervals for adjectives of age across baseline oral corpora and L3 Portuguese.	121
7.2	Linear regression results with 95% confidence intervals for adjectives of size across baseline oral corpora and L3 Portuguese.	123
7.3	Linear regression results with 95% confidence intervals for adjectives of physical appearance across baseline oral corpora and L3 Portuguese.	124
7.4	Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position.	126
7.5	Linear regression results with 95% confidence intervals for EM prepositional predicates across baseline oral corpora and L3 Portuguese.	129
7.6	Summary of word embedding results for baseline and L3 Portuguese corpora.	130
8.1	Token count across corpora and adjective type for adjectival predicates in copula structures.	133

LIST OF TABLES – *Continued*

8.2	Variance explained by logistic regression model. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.	134
8.3	Logistic regression results for ESTAR vs. SER with description adjectives across levels and baseline corpora.	135
8.4	Logistic regression results for ESTAR vs. SER with description adjectives across L1s and baseline corpora.	137
8.5	Variance explained by logistic regression model. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.	139
8.6	Logistic regression results for ESTAR vs. SER with evaluation adjectives across levels and baseline corpora.	139
8.7	Variance explained by logistic regression model for <i>estar</i> preference with verbal adjectives. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.	142
8.8	Logistic regression results for ESTAR vs. SER with verbal adjectives across levels and baseline corpora.	142
8.9	Logistic regression results for ESTAR vs. SER with verbal adjectives across L1 background and baseline corpora.	145
8.10	Total tokens and proportion of ESTAR tokens with EM preposition predicates across corpora.	145
8.11	Variance explained by logistic regression model for <i>estar</i> preference with verbal adjectives. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.	147
8.12	Logistic regression results for ESTAR vs. SER with EM prepositional predicates across L3 Portuguese levels and baseline corpora.	148
8.13	Logistic regression results for ESTAR vs. SER with EM prepositional predicates across L1 background and baseline corpora.	150
8.14	Summary of logistic regression results across corpora.	152
8.15	Summary of logistic regression results across corpora.	152
9.1	Summary of word embedding results for the Spanish corpora.	155
9.2	Summary of logistic regression results for baseline corpora. All results are significantly different from a chance probability.	156
9.3	Summary of word embedding results for baseline and L3 Portuguese corpora.	160
9.4	Summary of logistic regression results across corpora.	161
9.5	Summary of logistic regression results across corpora.	161

ABSTRACT

Previous research on third language (L3) acquisition has shown that the source language for transfer to the L3 can be either an L1, an L2, or both (Bardel & Falk, 2007; Flynn et al., 2004; Rothman, 2014). It has been hypothesized that either typological similarities between languages previously acquired and the target language (Rothman, 2010), or language status (L1 vs. L2) of previous acquired languages (Bardel & Falk, 2007) determine cross-linguistic influence. This dissertation investigates the acquisition of copula structures in L3 Portuguese by Spanish-English three groups of adult bilinguals: L1 English L2 Spanish, L1 Spanish L2 English, and L1 Spanish/English (i.e., heritage speakers of Spanish for the purposes of this dissertation). Language use by both native speakers (L1 Spanish, L1 English, and L1 Portuguese) and learners (L3 Portuguese) is analyzed using word embeddings and logistic regression modeling. The goal of these methods is to reveal patterns of copula use and acquisition. Copula constructions were chosen because they allow for the combined investigation of form, syntactic frame, and concept/meaning, as proposed by third language acquisition scholars. The main goal of this dissertation is to shed light on both transfer patterns from previously acquired languages (i.e., Spanish and English) on L3 Portuguese, and establish L3 Portuguese developmental patterns across bilingual groups. Results show evidence of L3 Portuguese development for all three groups of Spanish-English bilinguals. However, transfer patterns from Spanish and English onto L3 Portuguese are not the same across all groups, varying in degree depending on the copula construction. These results conflict with the *Typological Primacy Model*, which predicts that L3 acquisition in adulthood starts off from a wholesale transfer of the pre-acquired language system that is most typologically similar to the target language (Rothman, 2014). This dissertation offers support instead to L3 acquisition models that take into consideration structural characteristics of individual constructions, and how similar or different these are between source and target languages, including models such as the *Parasitic Model* (Hall et al., 2009).

CHAPTER 1

INTRODUCTION

The acquisition of a third language (i.e., L3 or any language acquired in adulthood after two or more languages have been previously acquired) from a cognitive approach is affected by previously learned languages, with the source language for transfer being either an L1, an L2, or both (Bardel & Falk, 2007; De Angelis, 2007; Flynn, Foley, & Vinnitskaya, 2004; Rothman, 2014; Slabakova & Pilar García Mayo, 2017). Some scholars argue that only one previously acquired language, whichever language is more typologically similar to the target language, is used as the source of wholesale transfer at initial stages of L3 acquisition (Giancaspro, Halloran, & Iverson, 2015; Rothman, 2010, 2014). Others argue instead for a property-by-property transfer from all previously acquired languages, with initial form-frame-concept connections of individual language constructions built based on cognate status and later revision or strengthening of these initial connections taking place throughout L3 development (Ecke, 2015; Ecke & Hall, 2014; Hall et al., 2009). All scholars in L3 acquisition, however, agree that much more research is needed to shed light on these issues, including the factors affecting language sources of initial transfer (e.g., typological similarity, L1 vs. L2 status) and cross-linguistic influence during L3 development (De Angelis, 2007; Slabakova & Pilar García Mayo, 2017).

Following the cognitive approach of the models mentioned above, this dissertation takes a cognitive perspective to L3 acquisition. The target language (i.e., L3 or the language that is being learned) is Portuguese, and the previously acquired languages are English and Spanish. A growing number of L3 Portuguese studies have been published in the last decade (Maimone, 2017), and, in many of these, Spanish and English represent the L1s and/or L2s. This language grouping is of interest because Spanish and Portuguese are sister languages from the Romance family (Montrul, Dias, & Santos, 2010), and although English “displays many influences from Romance languages at the lexical level” it shares morphosyntactic similarities

with Germanic languages (Falk & Bardel, 2010, p. 188). Also, these are the languages spoken across different countries in the Americas and it is not uncommon to find Spanish-English bilinguals in the United States. Spanish is also a salient language at University of Arizona, which has recently earned the designation of Hispanic-Serving Institution from the U.S. Department of Education. Understanding how Spanish-English bilinguals learn an additional language is relevant not only to the teaching of Portuguese as a foreign language but also to other language learning, especially Romance languages, such as French and Italian.

Although there is a considerable body of literature on L3 Portuguese acquisition by Spanish-English bilinguals (Child, 2014; Ionin, Montrul, & Santos, 2011; Iverson, 2009; Maimone, 2017), most of these studies make use of experimental data, with language perception and production elicited under controlled conditions for that given experiment. This dissertation takes a different approach by using corpus data, compiled from language produced for purposes other than this study. The participants that produced the language analyzed in this dissertation are L3 Portuguese learners who fall under one of the following bilingual groups:

1. L1 English L2 Spanish
2. L1 Spanish L2 English
3. L1 Spanish/English (i.e., heritage speakers of Spanish)

For the purposes of this dissertation, heritage speakers of Spanish are those who speak Spanish at home, but were in schools in the US where the primary language was English. In addition to the learner language produced by the groups above, I will analyze monolingual and bilingual native (L1) copula production in both English and Spanish, to establish patterns of copula use. The chosen structure is copula due to its ubiquity in language production and language teaching materials. I take a usage-based approach to investigate language use and acquisition in this dissertation, where “the psycholinguistic units with which people operate are identified through observation of their language use” (Tomasello, 2000, p. 62). More specifically, I explore how language use and acquisition take place by analyzing large amounts of linguistic data (both native and learner production) quantitatively. For that, I make use of corpus linguistics and computational linguistics methods to understand language developmental processes (i.e., how language features become more target-like over time).

1.1 Research Questions

My research questions reflect the need to investigate initial transfer (i.e., cross-linguistic influence at lower levels of proficiency) separately from L3 development (i.e., differences in patterns of language use across proficiency levels). My main goal is to identify underlying patterns of copula use in L3 Portuguese produced Spanish-English speakers to shed light on not only L3 Portuguese development but also cross-linguistic influence. I also include in my investigation patterns of copula use in Spanish produced by monolingual and bilingual speakers, and English produced by monolingual and bilingual speakers. My research questions are then as follows:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?
2. How do patterns of L3 Portuguese production by Spanish-English bilinguals compare to the source languages (i.e., Spanish and/or English) versus the target language (i.e., Portuguese) at each of the three levels of L3 proficiency?
 - a. Which language is the source of initial transfer for each Spanish-English bilingual group?
 - b. At what point (i.e., first, second or third semester) do L3 Portuguese patterns of copula use become most similar to L1 Portuguese (as opposed to most similar to L1 Spanish or L1 English)?
 - c. How similar are the L3 development paths across the three Spanish-English bilingual groups?

1.2 Overview

There are eight chapters in this dissertation in addition to this introduction: literature review, methods, word embedding results for Spanish baseline corpora, word embedding results for baseline corpora, logistic regression results for baseline corpora, word embedding results for

L3 Portuguese, logistic regression results for L3 Portuguese, and Discussion. The literature review covers research on L3 acquisition in a broad sense, including the main theoretical methods currently being discussed. It also describes what we know about copula structures in the three languages at hand. The third part of the literature review approaches what is known about copula acquisition in adults and children.

The methods chapter builds on a description of the research design. This chapter also presents corpora description. Finally, this chapter ends with descriptions for the two quantitative methods of linguistic analysis used in this dissertation (i.e., word embeddings and logistic regression).

The first results chapter addresses the first research question for Spanish only. The main goal of this chapter is to establish the Spanish baseline for Spanish transfer in later results chapter. The second and third results chapters also answer the first research question, with comparisons between potential transfer languages (i.e. Spanish and English) and the target language (i.e, Portuguese). The last two results chapters discuss results for the L3 Portuguese data, answering the second research question. The final chapter is a discussion of the findings, and it includes implications, limitations, and future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This literature review is divided into three parts. First I discuss L3 acquisition, including the main L3 acquisition models currently being debated in the literature. I also review the literature on L3 Portuguese acquisition. The second part is a discussion of copula structures in the three languages at hand (i.e., English, Portuguese, Spanish). The last and third part of this literature review examines what is known about *ser* and *estar* copula acquisition in adults and children.

2.2 L3 Acquisition

Although multilingualism is a common phenomenon around the world, most research on second language acquisition either assumes language learners are monolingual or ignores language background completely (De Angelis, 2007). The field of third language acquisition research has tried for over a decade to address the complex processes involved in the acquisition of an additional non-native language when two or more languages have been previously acquired (Alonso & Rothman, 2016; Cabrelli Amaro, 2017; Carvalho & Bacelar Da Silva, 2006; Child, 2014; De Angelis, 2007). It is important to note that third language acquisition researchers do not see bilinguals as two monolinguals in terms of cognition and thus acknowledge the complexity of cross-linguistic influence (De Angelis, 2007; Forsyth, 2014). Because of this perspective, scholars argue that third language acquisition research can add to more broad language acquisition research since “looking at multilingual transfer patterns permits a unique window into language and cognition in ways that cannot be seen in monolingualism or bilingualism” (Rothman, 2014, p. 181).

Most L3 acquisition research thus far has focused on lexis and phonology, with the former being most productive in the literature (Forsyth, 2014). Until recently, many scholars were actually skeptical that morphology and syntax could be transferred from a non-native language to another (De Angelis, 2007; Forsyth, 2014). The studies reviewed here contradict this no-transfer assumption by showing syntactic/morphosyntactic transfer of Differential Object Marking (DOM) (Giancaspro et al., 2015) and of object clitic pronouns (Montrul et al., 2010). Subjunctive mood transfer is also addressed, by two studies (Carvalho & Bacelar Da Silva, 2006; Child, 2014).

2.2.1 L3 Acquisition Models

In the next few sections I will describe the main models of cross-linguistic influence on L3 acquisition, in addition to other more recent or not as widely addressed L3A models. These models take into account factors such as language status (L1 vs L2) (Bardel & Falk, 2007; Schwartz & Sprouse, 1996), and typological similarities between transfer and target language (Ecke, 2015; Ecke & Hall, 2014; Hall et al., 2009; Rothman, 2010). One of the goals of this dissertation is to verify which of these models fit the evidence found in the learner data.

2.2.1.1 L1 Transfer Hypothesis

Early second language acquisition research assumed L1 cross-linguistic influence took place (Bohnacker, 2006; De Angelis, 2007), with some scholars arguing for a Full Transfer/Full Access model where the syntax for an additional language to be learned is built from the initial transfer of the entire L1 syntax (Schwartz & Sprouse, 1996). To support this model, a large body of research on verb placement (e.g., Subject-Verb-Object or V1, X-Verb-Subject or V2) in L2/L3 German was initiated. Schwartz & Sprouse (1994) analyzed the verb placement in L2 German by an adult speaker of Turkish, a language that does not implement the verb-second phenomenon. The participant's 26-month L2 German spoken production was separated into three stages of acquisition, with results showing that his L2 German grammar at the last stage still differed from L1 German syntax, i.e. the learner was still not fully

realizing V2 structures. The authors argue that a person's L1 sets certain Universal Grammar (UG) parameters that learners of additional languages are unable to overcome (Schwartz & Sprouse, 1994, 1996). As a consequence, the final stage of an L2 will never have the same cognitive status as the final stage of an L1, because those UG parameters are only set during the acquisition of a first language (Schwartz & Sprouse, 1996).

Håkansson, Pienemann, & Sayehli (2002) refute the Transfer/Full Access hypothesis, and argue that even when a certain structure is present in both L1 and L2, learners are not necessarily able to transfer/produce that structure in an additional language. The authors collected L3 German data from spoken interviews cross-sectionally (10 participants in their first year of German, 10 in their second year). All participants were L1 Swedish speakers, who had previously acquired English as their L2. Although German and Swedish are not mutually intelligible, both languages realize a verb-second position for affirmative main clauses. Results show that learners in their first year of L3 German start off by producing the Subject-Verb-Object word (which is non-target-like) exclusively, and at later stages they produce a non-target-like intermediate structure (Adverb-NPsubject-Verb-X) that is also ungrammatical in Swedish (their L1). Although the authors acknowledge that English could have played some role in the German acquisition of these Swedish speakers, they oppose the idea that syntactic transfer could occur from an L2. The authors also argue that no existing L3 acquisition model would take the effect of English (an L2 in this study) into account (this study by Håkansson et al. (2002) precedes the current L3 acquisition models).

To add to the argument that an L3 acquisition model that takes both L1s and L2s into account is needed, Bohnacker (2006) also investigated verb placement in the L2/L3 German production of two groups of Swedish speakers: three L3 German learners with prior knowledge of English and three L2 German learners with very limited knowledge of English. The author used an oral elicitation task after a four-month long German course. Results show that L2 German learners were able to produce the V2 order in German accurately in 100% of obligatory contexts, while the L3 German learners had an accuracy rate of 50%, revealing a detrimental influence of English.

As seen in the studies above, there is no consensus in the literature on whether an L1 has

a privileged role (over an L2) in cross-linguistic influence (Slabakova, 2016). Subsequent models of L3 acquisition have proposed different ways to untangle L1 and L2 cross-linguistic influence.

2.2.1.2 The Cumulative Enhancement Model

First proposed by Flynn et al. (2004), the *Cumulative Enhancement Model* assumes there is no wholesale transfer of one mental grammar to the L3 interlanguage. Instead, property-by-property facilitative transfer happens from both the L1s and L2s throughout L3 development (Alonso & Rothman, 2016; Falk & Bardel, 2010; Green, 2017).

Arguing that the L1 does not play a privileged role in L3 acquisition, Flynn et al. (2004) studied the acquisition of restrictive relative clauses in L3 English by speakers of L1 Kazakh and L2 Russian. Participants comprised of 33 adults and 30 children (ages 10-11). The authors used an elicited imitation task, where participants were read out a sentence containing a relative clause and were then to repeat it as exactly as possible. Participants' responses were correct if they matched the stimulus sentence, and incorrect when they did not match. Results indicate that the adult participants performed better at the experimental task than the children, which the authors interpret as evidence that a fully-acquired L2 contributes to L3 acquisition (in addition to L1 transfer).

The main criticisms of this model are that 1) it does not explain errors caused by cross-linguistic influence, and 2) it confounds transfer and L3 development since “what appears to be ‘facilitative transfer’ into the L3 may simply be acquisition” (Giancaspro et al., 2015, p. 192). Critics of this model also argue that there has not been a lot of empirical evidence “that supports the [Cumulative Enhancement Model] unambiguously” (Rothman, 2014, p. 183).

2.2.1.3 L2 Status Factor

First proposed by Bardel & Falk (2007), this model of L3 acquisition argues that due to the way it is represented in the brain, an L2 (i.e., a non-native language, probably learned during

adulthood) is the only source of initial transfer and cross-linguistic influence throughout L3 development (Cabrelli Amaro, Amaro, & Rothman, 2015; Falk & Bardel, 2010; Green, 2017).

Arguing for full-fledged transfer from an L2, and refuting the idea that only lexical transfer occurs from an L2 – as argued by Håkansson et al. (2002) – Bardel & Falk (2007) investigated the placement of negation in L3 Swedish and L3 Dutch. Their nine participants were distributed over a broad combination of L1/L2 backgrounds: L1 Dutch L2 English L3 Swedish, L1 English L2 German L2 Dutch L3 Swedish, L1 Hungarian L2 Dutch L3 Swedish, L1 Albanian L2 German L3 Dutch, L1 Italian L2 German L2 Dutch L3 Dutch, L1 Swedish L2 English L3 Dutch. Bardel & Falk (2007) attempt to disentangle this large number of L1s and L2s and argue that their results show syntactic transfer (i.e., negation placement) always occurs from an L2, and that the L2 status factor is stronger than typological similarity between an L1 and the L3.

The L2 Status Factor model proposes that cross-linguistic influence can be whole or partial, but the L2 is the exclusive transfer source (Alonso & Rothman, 2016; Giancaspro et al., 2015). This model is based on the “Foreign Language” effect, which argues that L1s are suppressed during L3 acquisition and production because they are not foreign, and the learner wants to avoid “sounding” as a speaker of their native language (Falk & Bardel, 2010; Forsyth, 2014). The main criticism of this model is that there has been evidence supporting transfer from L1s, especially when the L1 is more typologically similar to the L3 than the L2 is (Giancaspro et al., 2015).

2.2.1.4 The Typological Primacy Model

Initially proposed by Rothman (2010), the *Typological Primacy Model* argues strongly for the idea of wholesale initial transfer from either an L1 or an L2 - whichever language shares more typological similarities with the L3 (Alonso & Rothman, 2016; Cabrelli Amaro, 2017; Cabrelli Amaro et al., 2015). These similarities are assessed based on the following linguistic cues, in this order of importance: 1) lexicon, 2) phonetics/phonology, 3) morphology, and 4) syntax (Slabakova & Pilar García Mayo, 2017). After initial transfer, there is only L3

development (Alonso & Rothman, 2016), i.e., the L3 interlanguage uses a previously acquired language as a jump-start model, and from that point on, the L3 develops through language learning/acquisition. It is important to highlight here that the proponents of the *Typological Primacy Model* believe typological similarity assessment between languages is implemented subconsciously by the linguistic parser in the learner's brain, not being influenced by the learner's perception of similarity (i.e., psychotypology) (Giancaspro et al., 2015; Rothman, 2014). In other words, this model "rejects the notion that conscious psychotypological assessment on the part of the learner brings anything to bear" (Rothman, 2014, p. 185).

Critics of the *Typological Primacy Model* argue that this model does not take into account other factors that go beyond typology such as feature frequency found in the input (Slabakova, 2012), and processing complexity of individual structures (Slabakova, 2016). Others argue that aiming at identifying "single source for [cross-linguistic influence] [...] might not be realistic in any case" (Ecke, 2015, p. 155). There is also evidence that the written mode, when it allows learners more time to produce language (as opposed to improvised speech, or online production), shows cross-linguistic influence from more than one language (Forsyth, 2014). However, researchers seem to agree that "there is stronger support for the *Typological Primacy Model* than for the other three models, suggesting that the model is generally on the right track" (Slabakova, 2016, p. 655).

There is a general agreement that "Spanish is overall more structurally similar to [Brazilian Portuguese]" (Cabrelli Amaro et al., 2015, p. 23) than to American English. Supporters of the *Typological Primacy Model* also argue that Spanish and Portuguese will be selected as the most typologically similar languages due to shared lexicons (Giancaspro et al., 2015, p. 193). The literature on L3 Portuguese acquisition by Spanish/English speakers that support this model is large.

Giancaspro et al. (2015) investigated Differential Object Marking (DOM), which is an overt Case marking found in Spanish (but not in English and Portuguese), where human direct objects are marked with the preposition **a** while non-human direct objects are unmarked (e.g., Veo **a** mi amiga/*I see my friend* vs. Veo \emptyset mi película/*I see my movie*). The authors made use of grammaticality judgement tasks in all three languages (i.e., English, Spanish,

and Portuguese) with three groups of Spanish-English bilinguals learning Portuguese as an L3: L1 Spanish L2 English, L1 English L2 Spanish, L1 English L1 Spanish. This latter group was comprised of heritage speakers of Spanish, a participant group commonly found in the United States. All participants were enrolled in first semester Portuguese at university level, and L2 Spanish speakers were all placed at high-intermediate to advanced proficiency level. Results show all three groups behaving similarly, which suggests all learners are transferring from Spanish when making grammaticality judgements in Portuguese, providing support for the *Typological Primacy Model*.

Linguistic features that constrain the use and positioning of object clitic pronouns in Spanish and in Portuguese are often similar to those that constrain the use of Differential Object Marking (DOM) in Spanish, such as object animacy and definiteness (Tagliamonte, 2012). Object clitic pronoun use and perception was studied by Montrul et al. (2010) through an oral production and an acceptability judgment task, with three groups of participants: L1 Spanish L2 English, L1 English L2 Spanish, and native speakers of Portuguese. Clitic pronouns were chosen as the linguistic variable due to the contrasts among the three languages. First, Spanish is a clitic doubling language (1). Although both Brazilian Portuguese and Spanish have the prepositions *para* and *a*, these are used differently in both languages, with *a* serving as Differential Object Marking (DOM) in Spanish (2). In Brazilian Portuguese both *para* and *a* can be used in dative constructions. Null objects are common in Portuguese and allowed in Spanish, under different constraints (i.e., restricted to inanimate and non-definite objects). There are two clitic positions in Spanish (i.e., proclisis with finite verbs, and enclisis with non-finite verbs), while in Brazilian Portuguese proclisis is the norm in spoken production. In addition, clitic climbing has disappeared in Brazilian Portuguese (3).

(1) Portuguese: Dei isto ao senhor. *I gave this to the gentleman.*

Spanish: **Le** di esto al señor. *(To him) I gave this to the gentleman.*

(2) Portuguese: Vejo o João. *I see John.*

Spanish: Veo **a** Juan. *I see (to) John.*

(3) Portuguese: Pedro vai se levantar. *Peter will (himself) get up.*

Spanish: Pedro se va a levantar. *Peter (himself) will get up.*

The errors produced in the oral task in Montrul et al. (2010)'s study were traced back to Spanish for the two bilingual groups. In the acceptability judgement task, the two learner groups behaved again very similarly, with L1 Spanish speakers behaving slightly more native-like. These results provide support for the *Typological Primacy Model*, since the authors conclude that there is initial transfer from Spanish for all learner groups.

Also related to syntactic cross-linguistic influence, Child (2014) used both a sentence completion task and a grammaticality judgment task to investigate mood (subjunctive) distinctions in Spanish and in Portuguese with three groups of bilinguals: L1 Spanish L2 English, L1 English L2 Spanish, L1 English L1 Spanish (i.e., heritage speakers of Spanish). Both Spanish and Portuguese have productive subjunctive forms, but in Portuguese all three forms (past, present, and future) are used, while in Spanish the future subjunctive is rarely used. The results of the tasks indicate that learners in all groups transfer the knowledge they have of mood distinctions in Spanish to Portuguese. An earlier study on subjunctive in the same context by Carvalho & Bacelar Da Silva (2006) also investigated present and future subjunctive use by second semester Portuguese learners. A comprehensive analysis of errors show that participants transfer heavily from Spanish. Both studies, then, offer evidence that support the *Typological Primacy Model*.

Focusing on morphosyntactic and lexical cross-linguistic influence, Maimone (2017) investigated the recognition and written production of L3 Portuguese morphosyntactic items (two past participle forms in Portuguese, regular and irregular, e.g., *tem aceitado*/(it) has been accepting vs. *foi aceito*/(it) was accepted), and cognate and non-cognate lexical items at different learning stages. Participants were first semester Portuguese learners (i.e., they had no prior exposure to Portuguese) divided into two groups: L1 English speakers with no knowledge of Spanish, and L1 English L2 Spanish speakers. Experimental tasks included a lexical decision task (i.e., selection of which of two words is in Portuguese), a multiple-choice recognition task (i.e., fill in the blank with the correct past participle form), and a controlled

production task (i.e., type in the lexical item or the past participle that completes a sentence). For the last two tasks, a translation in English was provided. Results show a very small influence of English cognates for the group with no knowledge of Spanish, but no English influence for the L2 Spanish group. For this last group, there was also a significant amount of transfer from Spanish. For the production and recognition of regular and irregular past participle forms, the group with no knowledge of Spanish outperformed the L2 Spanish group, which relied on their Spanish knowledge of participles. These findings support the *Typological Primacy Model*, since there seems to be (at times detrimental) cross-linguistic influence from Spanish only.

Also in the area of morphosyntax, Iverson (2009) used a translation and grammaticality judgment tasks to examine gender agreement knowledge and noun drop in Portuguese with three groups of participants: native speakers of Brazilian Portuguese, L1 English L1 Spanish learners of Portuguese, and L1 English L2 Spanish learners of Portuguese. The author found no differences among the three groups related to the production and perception of gender agreement and noun drop in Portuguese. Although absence of evidence is not evidence of absence, these findings do seem to support the *Typological Primacy Model*.

Coming from a slightly different approach, Ionin et al. (2011) examined a purely semantic phenomena: plural noun-phrase (NP) interpretation. In English, the use of a determiner in plural NPs selects “specific” for its meaning, while the lack of a determiner renders the NP as generic (4). In Spanish, however, the plural NP is usually preceded by a definite article (5), which allows for both interpretations. Brazilian Portuguese allows both bare plural NPs, which are interpreted as generic like in English, and also definite plural NPs (6), which allow generic and specific interpretations like in Spanish. The authors used an acceptability judgement task in all three languages with the following participant groups: monolingual native English speakers, monolingual native speakers of Spanish, monolingual native speakers of Brazilian Portuguese (these three groups were used as baseline), L1 English L2 Spanish speakers, and L1 English L2 Spanish speakers learning Portuguese as an L3. We are interested here in this last group of participants, whose results show “they were not fully target-like: allowed definite plurals for generic/kind readings, but they did not allow bare plurals for

generic/kind readings (especially at the lower proficiency levels)” (Ionin et al., 2011, p. 292). This finding supports the *Typological Primacy Model*, since the Portuguese learners seem to be transferring plural NP interpretation exclusively from Spanish. In a follow-up study with a lower number of L1 Spanish L2 English L3 BP learners, the authors achieve the same results (i.e., no differences in transfer between L1 Spanish and L2 Spanish speakers).

(4) **The** tigers eat meat. (specific)

Tigers eat meat. (generic)

(5) **Los** tigres comen carne. (generic and specific)

* Tigres comen carne.

(6) **Os** tigres comem carne. (generic and specific)

Tigres comem carne. (generic)

2.2.2 Other L3 Acquisition Models

As mentioned, a few additional third language acquisition models have been also advanced in the literature. Most of these additional models go against the idea that transfer is wholesale, and propose a property-by-property transfer that is not exclusively facilitative, but that can also explain errors in L3 production.

2.2.2.1 The Parasitic Model

This model attempts to explain lexical lapses and deviant production (i.e., errors) in the L3 caused by cross-linguistic influence (Ecke, 2015; Ecke & Hall, 2014; Hall et al., 2009). It proposes three stages for L3 lexical acquisition that includes three aspects of lexical production (i.e., form, syntactic frame, concept/meaning). The first two stages are related to early L3 word learning, where a form representation is first mentally linked to a cognate form in either an L1 or an L2 (hence the name *parasitic*, due to L3 form “attachment” to an L1 or L2

form representation in the mental lexicon) and later form-frame-concept connections (i.e., analyzed constructions that are more flexible) are built. The third stage is ongoing, and strengthens (and/or revises) these initial connections. From this perspective, errors in L3 lexical production occur because “until the word forms stabilize and emancipate themselves from host representations, they will be susceptible to [cross-linguistic influence]” (Ecke, 2015, p. 152).

Hall et al. (2009) studied the effect of cognate status in the acquisition of L3 German and L3 French verbs by L1 Spanish L2 English speakers. A total of 33 participants were presented with 27 German verbs, and 40 participants were presented 45 French verbs. All verbs were equally divided into 3 groups: 1) those that were cognate with the participants’ L1 (i.e., Spanish), 2) those that were cognate with the participants’ L2 (i.e., English), and 3) those that were not cognate with either Spanish or English. Each verb in the L3 was displayed for seven seconds and the presentation consisted of the L3 verb in the infinitive form, accompanied by the translated infinitive forms in both English and Spanish. Participants were tested immediately after the presentation phase, and a week later. The testing was a grammaticality judgment task, where participants were to choose between two sentences in their L3: one sentence was presented in the Spanish equivalent’s verbal frame, and the other in the English equivalent’s verbal frame. When the L3 verb is not a cognate with either the L1 or the L2, learners default to their L2 verb frame (i.e., English) for both L3 German and L3 French, favoring the L2 Status factor or foreign language effect. For cognate verbs, a greater effect was found for French verbs in Spanish frames than German verbs in English frames, which supports the parasitic cognate effect. The authors acknowledge that more studies, with different methods and different language combinations, are needed to confirm these results.

2.2.2.2 The Scalpel Model

Slabakova (2016) proposes the *Scalpel Model*, which includes different learning patterns for different language properties, depending on the structural characteristics and the frequency in the input of the language items being transferred. More details on how these patterns form

are yet to be explored, but the model attempts to explain cross-linguistic influence beyond initial stages.

The *Scalpel Model* is also an attempt to explain conflicting results from studies on the same group of L2/L3 English learners, more specifically L1 Basque L2 Spanish L3 English, L1 Spanish L2 Basque L3 English, L1 Spanish L2 English. These studies showed that learners were all able to reject null objects in English (e.g., English learners correctly identified constructions such as *I like ∅* as ungrammatical) (García Mayo & Slabakova, 2015) but were unable to detect acceptable topicalizations in English (e.g., *The salmon I haven't tried it yet* vs. **The salmon I haven't tried yet*) (Slabakova & García Mayo, 2015). Both null objects and topicalizations are much more frequent in Basque and Spanish than in English, yet the pattern of acquisition for these two structures is different, which indicates other factors at play (Slabakova, 2016).

2.2.3 Initial Transfer vs. L3 Development

Some scholars argue that initial transfer and L3 learning or development are two distinct processes (Green, 2017), and thus we should focus on perception and production of naive or true beginner learners because what “happens beyond the initial representations of L3 interlanguage grammar [...] is a question of L3 acquisition itself” (Alonso & Rothman, 2016, p. 687). Others argue for going beyond the initial state, to get a better sense of the entire learning process (Slabakova, 2016; Slabakova & Pilar García Mayo, 2017). There is indeed evidence that initial transfer can affect language production even at later stages of L3 acquisition, and that that influence can be attested if still present at these later stages (De Angelis, 2007). On the other hand, if at higher levels of proficiency an expected influence is not attested, that does not mean that it was not present at initial stages (Slabakova & Pilar García Mayo, 2017).

It is important to reiterate that some L3 acquisition scholars argue that transfer from a previously acquired language occurs only at initial stages. *The Typological Primacy Model*, for example, predicts that L3 acquisition in adulthood starts off from a wholesale transfer of a

pre-acquired language system, with no additional transfer during L3 development (Rothman, 2010). From this perspective, if any cross-linguistic influence is attested at higher levels of proficiency “it must have been there from the beginning of the acquisition process” (Slabakova & Pilar García Mayo, 2017, p. 69).

2.2.4 Summary of L3 Acquisition Models

Here is a summary of the five main L3 Acquisition models addressed in this literature review, and their main predictions.

- **L1 Transfer Hypothesis** (Bohnacker, 2006; Håkansson, Pienemann, & Sayehli, 2002; Schwartz & Sprouse, 1994, 1996): the syntax for an additional language to be learned is built from the initial transfer of the entire L1 syntax
- **L2 Status Factor** (Bardel & Falk, 2007; Falk & Bardel, 2010): an L2 (i.e., a non-native language, probably learned during adulthood) is the only source of initial transfer and cross-linguistic influence throughout L3 development
- **The Cumulative Enhancement Model** (Flynn et al., 2004): facilitative transfer from both L1 and L2 occurs onto the L3
- **The Typological Primacy Model** (Rothman, 2010): wholesale initial transfer is from either an L1 or an L2 - whichever language shares more typological similarities with the L3
- **The Parasitic Model** (Ecke, 2015; Ecke & Hall, 2014; Hall et al., 2009): no wholesale transfer, L3 lexical representations are “attached” to either L1 or L2 lexical representations, whichever form is more similar to the L3 form
- **The Scalpel Model** (Slabakova, 2016): no wholesale transfer, many factors play a role on how individual language constructions are transferred from previously acquired languages

2.3 Copula Structure

This dissertation focuses on copula constructions due to their ubiquity in language production and language teaching materials – one of the first structures in most language textbooks in any language is the verb to be (or *ser/estar*). Copula constructions are also of interest because they combine a number of lexical, syntactic and semantic components that have been rarely explored in conjunction thus far. These constructions allow for the combined investigation of form, syntactic frame, and concept/meaning, as proposed by Ecke (2015). This is relevant to this dissertation because “usage-based approaches to language learning in both L1 and L2 emphasize the interconnection between vocabulary and grammar” (Staples & Reppen, 2016, p. 18). In this section I briefly discuss this target feature in all three languages addressed in this dissertation (i.e. English, Portuguese, and Spanish) to establish overlapping and contrasting use of copula, which in turn brings to light possible facilitative and detrimental cross-linguistic influence.

While copula structures allow for the study of third language acquisition due to their complexity, the actual term, “copula” has a number of definitions across linguistics. Moro (2000) defines copula constructions as structures in which the verb *be* is the main verb (as opposed to *be* being an auxiliary verb as in *The girl is running*). As such, the following are examples of copula use, from Moro (2000).

- (7) A picture of the wall **is** impressive.
- (8) A picture of the wall **is** in the room.
- (9) A picture of the wall **is** the cause of the riot.
- (10) There **is** a picture of the wall in the room.

As seen in the examples 7-9 above (10 is a special case), the copula verb connects a subject with a nonverbal predicate (Moro, 2000; Moura, 2016; Sibaldo, 2011). The predicate of a copula verb can be an adjective phrase (7), a noun phrase (10), an adverbial phrase (11), or a prepositional phrase (8).

(11) He **is** here.

The same is true for Portuguese, the copula verb connects a subject with a nonverbal predicate. Here are some examples in Portuguese from Sibaldo (2011) and Moura (2016) that make use of copula verbs to connect a subject with a predicate:

(12) O trabalho **está** [pronto].

The paper is [ready].

(13) Eu **estou** [ferido].

I am [hurt].

(14) O João **é** [médico].

John is [a medical doctor].

(15) Eu **sou** [seu amigo].

I am [your friend].

(16) A criança **está** [no jardim].

The child is [in the garden].

(17) A criança **está** [longe de casa].

The child is [far from home].

The Spanish copula has been extensively studied from different theoretical perspectives and using a variety of methods (Woolsey, 2008). Spanish, as Portuguese (12-17), has two copula verbs (i.e., *ser* and *estar*) that encode all the meanings of the copula *be* in English (Finnemann, 1990; Garavito & Valenzuela, 2006; Geeslin & Guijarro-Fuentes, 2006; VanPatten, 2010). Both in Portuguese and in Spanish, the verb *ser* is considered unmarked or underspecified for temporal/aspectual information, while *estar* indicates a state (i.e., property *P* holds at a certain point *t* in time) (Finnemann, 1990; Garavito & Valenzuela, 2006; Schmitt, 1992;

Schmitt, Holtheuer, & Miller, 2004).

VanPatten (2010) summarizes the use of these two copula verbs in Spanish as follows:

A. Only *ser* is used with predicate noun phrases (NPs) (*Juan es estudiante*, i.e., John is a student) and with predicate prepositional phrases (PPs) to express origin (*Juan es de México*, i.e., John is from Mexico).

B. Only *estar* is used with locatives (*Juan está en casa*, i.e., John is at home).

C. Both *estar* and *ser* are used with adjectives, but not interchangeably (i.e., the meaning changes depending on the verb used): *Juan es triste* (i.e., John is always sad, as in a permanent state of sadness), *Juan está triste* (i.e. John is sad right now, as in a temporary state of sadness).

Compared to the abundance of research in Spanish, there have been fewer investigations into *ser* and *estar* as copula verbs in Portuguese (Sibaldo, 2011). However, two out of the three points above (i.e., A and C) are also true for Portuguese. Regarding B above for Portuguese, both *estar* and *ser* are used with spatial locatives. Thus, here is the summary of *ser* and *estar* copula verbs in Portuguese:

A. Only *ser* is used with predicate noun phrases (NPs) (*João é estudante*, i.e., John is a student) and with predicate prepositional phrases (PPs) to express origin (*João é do Brasil*, i.e., John is from Mexico).

B. Both *estar* and *ser* are used with locatives. The copula *estar* is used with mobile referents (*João está em casa*, i.e., John is at home), while *ser* is used with stationary referents (*A casa é no centr.*, i.e., The house is downtown)

C. Both *estar* and *ser* are used with adjectives, but not interchangeably (i.e., the meaning changes depending on the verb used): *João é triste* (i.e., John is always sad, as in a permanent state of sadness), *João está triste* (i.e. John is sad right now, as in a temporary state of sadness).

As pointed out in C, *ser* and *estar* are used with adjectival predicates with choice of verb

depending on speaker intent (Woolsey, 2008). Trying to get at the speaker intent is not an easy task, even when considering the context. This makes the study of the acquisition of the different meanings of these two copula verbs difficult (Woolsey, 2008). Errors here are also harder to get at, since incorrect use of *ser* instead of *estar* or vice versa would depend on the learner's communication intent.

Although *ser* and *estar* in Portuguese are used in similar contexts in Spanish, their uses do not match perfectly (Moura, 2016). For locatives, for example (point C above), *estar* is used in some contexts where in Portuguese we use *ser*, or even a third verb, *ficar*, e.g. Spanish: *¿Dónde está el baño?* (i.e, Where is the bathroom?), Portuguese: *Onde é/fica o banheiro?*. Portuguese presents this third copula verb, *ficar*, that can also indicate a change of state (Ribeiro, 2004), which is similar to copula *get* in English and copula *ponerse* in Spanish.

To address where copula use overlaps and diverges, the next sections provide a contrastive analysis across the three languages, divided into the different types of predicate.

2.3.1 Nominal Predicate

Nominal copula (i.e., NP V NP) is also called identity statements (Schmitt & Miller, 2007). In Portuguese and Spanish, identity copula structures, due to their permanent status, select *ser* as the copula verb (Moura, 2016; Schmitt & Miller, 2007). Examples from Moura (2016) and Soschen (2002):

(18) Maria **é** a mulher de Pedro.

*Maria **está** a mulher de Pedro.

Mary **is** the wife of Pedro (*Mary is Pedro's wife*).

(19) Carmen **es** la señora de Garcia.

*Carmen **está** la señora de Garcia.

Carmen **is** the wife of Garcia (*Carmen is Garcia's wife*).

These structures have been extensively studied across different languages, with evidence that the NP in the predicate slot can behave as a preverbal subject (Moro, 2000). This is reflected in Brazilian Portuguese in the number agreement of the copula verb, which can agree with either the subject or the predicate (Sibaldo, 2011). In non-null-subject languages, such as English, that is not the case, with the verb agreeing always with the subject. Although Spanish is also a null-subject language like Portuguese, an informal grammatically judgement test with a few Spanish speakers revealed that this varying agreement phenomenon does not occur in Spanish. Here are examples of identity copula structures in Portuguese with varying verb agreement from Sibaldo (2011).

(20) O João e a Maria **são** a causa da revolta.

John and Mary **are** the cause of the riot.

Juan y Maria **son** la causa de la revolución.

(21) O João e a Maria **é** a causa da revolta.

*John and Mary **is** the cause of the riot.

*Juan y Maria **es** la causa de la revolución.

(22) A causa da revolta **são** o João e a Maria.

*The cause of the riot **are** John and Mary.

*La causa de la revolución **son** Juan y Maria.

(23) A causa da revolta **é** o João e a Maria.

The cause of the riot **is** John and Mary.

La causa de la revolución **es** Juan y Maria.

This varying agreement phenomenon in Portuguese, however, is not very useful in the investigation of cross-linguistic influence, since the transfer choice from either Spanish or English will always be grammatical in Portuguese. Also, the use of the singular form of

ser preceded by a plural noun phrase is probably not very frequent in either monolingual Portuguese or L3 Portuguese production.

2.3.2 Adjectival Predicate

Broadly, *ser* and *estar* in both Spanish and Portuguese are semantically distinct in that *ser* is associated with properties that are permanent or inherent, while *estar* is characterized by acquired or temporary attributions (Schmitt & Miller, 2007; Sibaldo, 2011). While most adjectives can be used with both verbs (Geeslin & Guijarro-Fuentes, 2006; Schmitt et al., 2004; Woolsey, 2008), with change in meaning (24-25 below), some adjectives can only be used with either *ser* or *estar* (26-27) (Schmitt & Miller, 2007). Examples in Portuguese based on Sibaldo (2011):

(24) A Maria **está** muito feia.

María **está** muy fea.

Mary is very ugly.

(25) A Maria **é** muito feia.

María **es** muy fea.

Mary is very ugly.

(26) A Maria **está** grávida.

María **está** embarazada.

Mary is pregnant.

(27) *A Maria **é** grávida.

*María **es** embarazada.

Mary is pregnant.

This restriction, as shown in (26-27) above, is due to the lexical meaning of the predicate, e.g., pregnancy is never permanent. However, this is not always a straight forward relationship; it can be argued that adjectives such as *young* and *new* denote temporary properties, but these are used with *ser* only both in Portuguese and in Spanish, while the adjective *dead* can be seen as permanent, but it is used with *estar* (Schmitt et al., 2004; Schmitt & Miller, 2007).

This fixed copula use for some adjectives, however, is useful in collocation analyses. In other words, fixed copula constructions can be tested for transfer by looking at what adjectives these verbs occur with in Portuguese and in Spanish. In a corpus of native speaker production the adjective *pregnant* or *dead* in Spanish and in Portuguese will strongly collocate with the verb *estar* as opposed to *ser*. This may not be the case in a learner corpus, especially at lower levels of proficiency and in the case of no facilitative cross-linguistic influence.

Another useful fixed copula use is with past participle derived adjectives. In Spanish, only the copula *estar* can be used with an adjective derived from an accomplishment verb (e.g., *abierto* *open*, *cansado* *tired*, and *confundido* *confused*) (28-29) (Garavito & Valenzuela, 2006).

(28) A Maria **está** cansada.

María **está** cansada.

Mary ****is*** tired.

(29) *A Maria **é** cansada.

*María **es** cansada.

*Mary **is** tired.*

Some exceptions to the use of *estar* with the past participle of a verb in Portuguese include *casado* (i.e., married), *feito* (i.e., made), and *proibido* (i.e., prohibited). All of these require the use of *estar* in Spanish, while in Portuguese *ser* is required (30-32) (Carvalho & Bagno, 2015).

(30) A Milena **é** casada.

Milena **está** casada.

*Milena **is** married.*

(31) A casa **é** feita de madeira.

La casa **está** hecha de madera.

*The house **is** made of wood.*

(32) Deveria **ser** proibido estudar no domingo.

Debería **estar** prohibido estudiar los domingos.

*It should **be** prohibited to study on Sundays.*

In addition to the permanent vs. temporary attribution, another way to look at this choice is whether the speaker is making a reference to a general vs. an individual norm (Falk, 1979; Woolsey, 2008). Falk (1979) first proposed this distinction by arguing that saying that someone is beautiful or ugly in Spanish (33-34) is linked to whether the speaker is making a comparison to a general norm of all people (i.e., use of *ser* as in 33) or to an individual norm (i.e., use of *estar* as in 34).

(33) Juan **es** guapo.

*John **is** handsome (compared to all the other men, general norm).*

(34) Juan **está** guapo.

*John **is** handsome (compared to how he usually looks, individual norm).*

Regarding the adjectives that can be used with both *ser* and *estar*, speaker intent is at the heart of this choice, as seen above. In other words, “the choice between *ser* and *estar* has a pragmatic component” (Schmitt & Miller, 2007, p. 1910). Whenever syntax interfaces with other systems, such as pragmatics, attrition and incomplete learning tend to take place (Garavito & Valenzuela, 2006). As such, the use of *estar* compared to *ser* in these situations have been extensively studied in the acquisition of Spanish as an additional language (as

described later in this literature review) and to a lesser extent in first language acquisition (Schmitt & Miller, 2007).

The distribution of these adjectives across *ser* and *estar* in a corpus will depend on the topic of discourse (i.e., established by frame of reference), and as such, it is difficult to determine any cross-linguistic influence on this choice.

2.3.3 Prepositional Predicate

Similarly to the semantic restriction imposed by some adjectives, prepositional phrases in the predicate position of a copula verb never allow for interchangeable use of *ser* and *estar*. Examples again are based on Sibaldo (2011):

(35) A Denise **está** na praia. (location)

*A Denise **é** na praia.

Denise **está** en la playa.

Denise **is** at the beach.

(36) A Denise *está* com sua mãe. (company)

*A Denise **é** com sua mãe.

Denise **está** con su madre.

Denise **is** with her mom.

(37) O Marcos **está** com dinheiro. (possession)

*O Marcos **é** com dinheiro.

Marcos **tiene** dinero.

Marcos **has** money.

(38) O feijão **está** por cinco reais. (cost)

*O feijão **é** por cinco reais.

El frijol **está** por cinco reales.

Beans **are** five dollars.

(39) *A chave **está** de ouro.

A chave **é** de ouro. (material)

La clave **es** de oro.

The key **is** made of gold.

(40) *O dinheiro **está** para Denise.

O dinheiro **é** para Denise. (beneficiary)

El dinero **es** para Denise.

The money **is** for Denise.

(41) *Esse tapete **está** da China.

Esse tapete **é** da China. (origem)

Esta alfombra **es** de China.

This rug **is** from China.

(42) *A missa **está** de sete horas

A missa **é** de sete horas. (duration)

La misa **es** de siete horas.

The mass **is** seven hours long.

There are some fixed expressions in Portuguese that deviate from the uses above, e.g. *o Senhor é convosco* (*The Lord is with thee*). These are only found in very specific corpora, such as religious texts.

From the examples above, the use of *ser* with the preposition *com* to express temporary possession (37) or temporary status (e.g., *Estou com fome/I'm hungry*) is not present in Spanish. All the other cases can be directly translated to Spanish. Since there is no equivalent to this structure (i.e., *estar + com + NP*) in either Spanish or English, its use is an indication of L3 Portuguese acquisition (as opposed to transfer).

Another contrast of interest that has been previously mentioned is the use of *ser* with locatives that are stationary in Portuguese, which differs from Spanish, e.g. Spanish: *¿Dónde está el baño?*, Portuguese: *Onde é o banheiro?* (i.e., Where is the bathroom?).

2.3.4 Extension of *estar*

As discussed above, *estar* is marked for aspect. As such, the restrictions presented in the previous sections are, in some instances, the result of (individual and historical) language change (Garavito & Valenzuela, 2006; Geeslin & Guijarro-Fuentes, 2006). These processes at times extend the use of *estar* to contexts where previously *ser* was to be used. For example, adjectives of size, such as *pequeno*/small, are prescriptively used with *ser*, as in *el niño es pequeno*/the boy is small. However, in certain communities this construction at times is used with *estar* as in *el niño está pequeno*/the boy is small. In fact, there has been extensive research on the semantic extension of *estar* in L1 Spanish, especially in copula + adjective contexts in monolingual and bilingual communities (Bessett, 2015; Cortés-Torres, 2004; Geeslin & Guijarro-Fuentes, 2008; Salazar, 2007). Silva-Corvalán (1986) studied the extension of *estar*, also known as innovative *estar*, with adjectival predicate in an English-Spanish bilingual community in Los Angeles, United States. The results show that the lower the level of Proficiency in Spanish, the higher the incidence of innovative *estar*.

The extension of *estar* is not exclusive to bilinguals (Salazar, 2007). Focusing on monolingual Spanish in Cuernavaca, Mexico, a city 40 miles away from Mexico City, Cortés-Torres (2004)

found that 23% of copula instances presented the use of *estar* when *ser* was expected with adjectival predicates. Comparing both Spanish monolingual and Spanish-English bilingual communities, Bessett (2015) conducted 40 sociolinguistic interviews in Sonora, Mexico and Arizona, United States. The study reveals that innovative use of *estar* is used in 16.2% of copula + adjective instances by the speakers in Sonora (Mexico) and 20.8% by the speakers in Arizona, with no significant difference between the two groups. Adjectives of age (e.g., *joven/young*) and size (e.g., *grande/big*) are the two types of adjectives that show up as favoring the use of *estar* in both communities. Type of adjective was also the most important factor predicting the use of innovative *estar* in a Spanish-English bilingual community in New Mexico (Salazar, 2007). Two other significant predicting factors for innovative *estar* use are related to adverbs, namely the use of time adverbials (e.g., *ya/already*) and intensifiers (e.g., *muy/very*).

2.3.5 Verbs Other Than *ser* and *estar*

In addition to *to be*, *ser*, and *estar*, there are other copula verbs in Portuguese, Spanish and Portuguese. In Brazilian Portuguese, *ficar* is one of these additional copulas (Rebouças, 2018; Schmitt, 2013). In some instances *ficar* is marked for aspect, indicating either a change from a previous state (44, 49) (Rebouças, 2018) or a temporary state (46-47) as *estar* (Schmitt, 2013). The unmarked *ficar* copula is used with stationary locatives (45), as already mentioned (e.g., *Onde fica o banheiro?/Where's the bathroom?*).

(43) Pedro **está** cansado.

Pedro **is** tired.

(44) Pedro **ficou** cansado.

Pedro **got** tired.

The copula *ficar* is accompanied by the same types of predicates as *ser* and *estar*, including adjectival (44), prepositional (45-46, 49), and adverbial (47) predicates (Rebelo & Osório, 2006).

(45) A PUC **fica** na Gávea.

PUC **is** in Gávea.

(46) Elas **ficaram** com o Chico na Igreja.

They **were** with Chico at church.

(47) Eu e a Leila **ficamos** juntos alguns anos e depois separamos.

I and Leila **were** together for a few years and then we split up.

Depending on the predicate, *estar* means *become* (44, 49) or *stay* (46-47) (Schmitt, 2013).

(49) A Maria **ficou** com fome.

Mary **got** hungry.

There are four “commonly occurring Spanish verbs” (Bybee & Eddington, 2006, p. 323) that when used with an adjectival predicate express a change of state, similarly to *ficar* in Portuguese: *quedarse* (50), *ponerse* (51), *volverse* (53), and *hacerse* (52) (Bybee & Eddington, 2006; Wilson, 2010, 2014).

(50) Celeste se **ha quedado** ciega.

Celeste’s **gotten** blind.

(51) De repente se **pone** furiosa.

All of a sudden she **gets** furious.

(52) Se **hace** tarde.

It’s **getting** late.

(53) Se ha **vuelto** loco.

He’s **gone** crazy.

It is important to note that not all scholars agree that these verbs are actual copula, some classify them as semi-copula and others argue they are regular lexical verbs (Schmitt, 2013).

2.4 Copula Acquisition

Copula acquisition in L3 Portuguese by English-Spanish bilinguals pose an opportunity to disentangle cross-linguistic influence, due to the cognate copulas in Spanish (i.e., *ser* and *estar*) and the lack of marked/unmarked duality in English (i.e., *be* is the only copula verb that maps to *ser* and *estar* in Portuguese). As described in the previous section, *ser* and *estar* present similar semantic and syntactic frames (but not completely overlapping) across Spanish and Portuguese, and thus could potentially be directly transferred from Spanish.

There has been extensive research on L2 Spanish copula verb acquisition (Garavito & Valenzuela, 2006; Geeslin & Guijarro-Fuentes, 2006; Ryan & Lafford, 1992), which is not the case for L2 Portuguese acquisition. In the next sections, copula acquisition in Spanish by either Portuguese-speaking or English-speaking adults is discussed.

2.4.1 Copula Acquisition in Spanish

As mentioned, Spanish copula acquisition by English speakers has been broadly studied (Garavito & Valenzuela, 2006; Guntermann, 1992; Ryan & Lafford, 1992; VanPatten, 1987). From this body of research, the main stages of acquisition for *ser* and *estar* copula verbs in Spanish in adults seem to be the following (Ryan & Lafford, 1992; VanPatten, 1985, 1987):

- Stage 1: learners do not use a verb in copula contexts: Juan muy inteligente (i.e., John very smart).
- Stage 2: learners use only *ser* in all contexts where a copula verb is needed.
- Stage 3: learners introduce *estar* as an auxiliary verb in the continuous aspect: Juan est'a estudiando (i.e., John is studying).
- Stage 4: learners overuse *estar*.
- Stage 5a: learners start using *estar* with locatives (e.g., Juan está aqui, 'John is here').

- Stage 5b: learners use *estar* as a copula feature with adjectives, showing the ability to differentiate its meaning from *ser*.

There is conflicting evidence on whether *Stage 5a* precedes or follows *Stage 5b* (Guntermann, 1992; Ryan & Lafford, 1992; VanPatten, 1987). Although later studies seem to confirm the stages proposed by VanPatten (1987), which places the use of *estar* with locatives before the use of *estar* as with adjectives (Woolsey, 2008). Regardless of the exact order, the use of both copula *ser* and *estar* with adjectival predicates showing a differentiation in meaning in Spanish is one of the last structures to be acquired by L2 learners (VanPatten, 2010). As such, this distinction has been extensively studied in the acquisition of Spanish as a second language.

Earlier studies focus on either distribution of use only (i.e., percentage of occurrences) or distribution of errors (Finnemann, 1990; Guntermann, 1992). Finnemann (1990) analyzed copula use in longitudinal speech data produced by three adult L2 Spanish learners (L1 English), sampled in regular periods over a six-month period. The data in Guntermann (1992) came from 20 oral examinations of Peace Corps volunteers (L1 English L2 Spanish). Results show not only that *estar* is less frequently used than *ser*, but also that when *estar* is used, more errors are made by the speakers. These results again confirm the stages proposed by VanPatten (1985), where *ser* is more prominently used, especially at lower levels of Spanish proficiency. The authors interpret these results by arguing that since *ser* is the unmarked choice, learners default to it even when the marked choice is needed.

More recent studies include the use of a variationist approach (Geeslin & Guijarro-Fuentes, 2006; Woolsey, 2008) and comparisons in copula choice including Spanish-English bilinguals (Garavito & Valenzuela, 2006). Woolsey (2008) studied the choice of adjective in the predicate slot of copula constructions in the language choice of L2 Spanish learners whose L1 was English. Confirming the stages first proposed by VanPatten (1985), results show a steady decline of *ser* frequency as proficiency level increases, with *estar* showing the opposite trend. As expected, the task type is the strongest predictor of *estar* use across all levels of proficiency in the regression modeling. The most relevant results from this study is that adjectives that are derived from verbs start to predict the use of *estar* in the regression starting at level 2,

with stronger prediction odds at higher levels.

With the purpose of comparing groups of Spanish-English bilinguals, Garavito & Valenzuela (2006) investigated the use of *ser* and *estar* with adjectival predicates in Spanish by giving Spanish-English bilinguals (L2 Spanish L1 English, and L1 Spanish L1 English) a grammaticality judgment task (i.e., minimal pair sentences with the choice of *ser* and *estar*), and a story completion task (multiple choice, with the minimal pair sentences again). A group of monolingual Spanish speakers was used for baseline comparisons. Results show that L2 Spanish speakers accept sentences that combine *ser* with verbal adjectives (e.g., *inacabado unfinished*), which are ungrammatical, much more often than the two other groups (i.e., L1 Spanish groups). The L2 Spanish group behaves the same way as the other two groups in all other instances of copula use with adjectival predicates. This suggests that incomplete acquisition is taking place with the L2 Spanish speakers when it comes to participle use in copula constructions.

The only study on L2 Spanish copula acquisition by Portuguese speakers (Geeslin & Guijarro-Fuentes, 2006) also employed a variationist approach due to the varying nature of copula choice for a given context. The authors argue that an analysis of learner errors in this situation would not be adequate. Instead, they use statistical models comparing features that are predictive of copula choice for native speakers compared against L2 Spanish learners. The data elicitation task for this study consisted of 28 narrative paragraphs to establish the discourse situation, with a binary copula choice question at the end of each prompt (e.g., Paula: ¿Por qué no te llevas bien con Alicia? *Why don't you get along with Alicia?* A) Raúl: Alicia no *está* muy feliz A) Raúl: Alicia no *es* muy feliz, *Alicia is not very happy*). Results show that L2 Spanish L1 Portuguese speakers aligned their frequency distribution of copula choices with L1 Portuguese speakers instead of L1 Spanish speakers. However, the language features that predict copula choice were similar across the L1 groups, with the L2 Spanish group displaying added predictive factors (i.e., animacy and experience with the referent). The authors interpret these differences as possible incomplete acquisition of Spanish. In addition, the regression models for the L1 data (e.g., L1 Spanish and L1 Portuguese) were able to explain much more of the variation (Nagelkerke R^2 of .86 and .85 respectively) compared

to the model for the L2 data (R^2 of .59), which means there was much more variability in the L2 data.

2.5 Conclusion

As it can be seen in the L3 acquisition studies reviewed in this section, linguistic structures in previously acquired languages and the L3 may overlap or contrast, with overlapping features being acquired faster and more accurately than contrasting properties, which are acquired at a slower pace (Slabakova & Pilar García Mayo, 2017). That means that contrasting features have more opportunities to cause more errors. However, while some linguistic variables are better studied in learners' errors, others are better investigated using other methods (e.g., distribution of collocations in the production of learners across different proficiency levels). Divergent word order production (e.g., Subject-Verb-Object vs. X-Verb-Subject), for example, is harder to acquire because they may reflect discourse appropriateness that does not always cause major communication problems (Slabakova & Pilar García Mayo, 2017).

When it comes to *ser* and *estar* in Portuguese, this chapter has discussed both obligatory contexts where one or the other is used and the matter of choice depending on the intended meaning. In both situations, there are overlaps and contrasts in the use of *ser* and *estar* between Portuguese and Spanish. This dissertation aims at first confirming these contrasts and overlaps in the use of *ser* and *estar* on baseline corpora (in monolingual Portuguese and Spanish). The monolingual English corpora will be used in cognate status analyses for form similarities between English and Portuguese adjectives. The second step is to extract patterns of *ser* and *estar* use in the L3 Portuguese corpus, to shed light on cross-linguistic influence of Spanish and English (based on the baseline corpora findings) across the three levels of proficiency that the L3 data is divided into.

Another point to highlight based on this literature review is that most L3 Portuguese acquisition studies use acceptability judgment tasks (also called grammaticality or preference judgement tasks, depending on the linguistic variable) for perception studies and some sort of elicitation task for investigation of L3 production (including error analysis). The L3/L2

German/Dutch/Swedish acquisition studies reviewed in this chapter do collect what they frame as “naturalistic” oral production (Bardel & Falk, 2007; Bohnacker, 2006; Håkansson et al., 2002; Schwartz & Sprouse, 1994, 1996), but the number of participants in each of these studies do not reach a total of ten participants. Corpora comprised of learner data from a large number of participants, with language produced for reasons other than specific studies are rarely used in L3 acquisition research. The data used in this dissertation, by contrast, is comprised of written assignments by L3 Portuguese learners.

CHAPTER 3

METHODS

3.1 Introduction

As seen in the literature review, most studies on L2 Spanish copula verb acquisition have relied on frequency distributions. In these studies, higher frequency of *estar* across proficiency levels is seen as evidence of its acquisition (Finnemann, 1990; Guntermann, 1992). Notable exceptions to the frequency-based studies are the only study on L2 Spanish copula acquisition by Portuguese speakers (Geeslin & Guijarro-Fuentes, 2006) and a study with English speakers (Woolsey, 2008), in which the authors made use of use statistical models, namely logistic regression.

Regarding L3 Portuguese Acquisition by Spanish-English Bilinguals, most scholars opted for data elicitation in experimental settings (Child, 2014; Ionin et al., 2011; Iverson, 2009; Maimone, 2017). I will take, however, a different approach by using corpora of language produced for purposes other than this study. I will also implement quantitative analyses of copula use, instead of relying on individual lexeme frequencies, to take into account the context of each token (i.e., each occurrence of copular *ser* and *estar*). This dissertation consequently fits into what is called a Type A design in Corpus Linguistics, taking as the unit of analysis each occurrence of a linguistic feature in a variationist approach (Biber & Jones, 2009), which makes use of regression to establish patterns of language use. Each of the quantitative methods used in this dissertation is meant to reveal patterns of copula use, e.g., what words co-occur more frequently with each copula verb, or what predicate types or tense/aspect/mood favor certain copula verbs. Patterns of use refer then to any systematic variation (i.e., variation that can be accounted through statistical methods) present in the corpus data. The quantitative methods chosen are provide a complementary view of copula use across different groups of speakers.

In this chapter I describe the corpora used, which are divided into baseline corpora (i.e., L1 monolingual language produced by Spanish and Portuguese speakers) and a learner corpus. This last corpus consists of written L3 Portuguese produced by three groups of Spanish-English bilinguals: 1) L1 English L2 Spanish, 2) L1 Spanish L2 English, and 3) L1 Spanish/English (i.e., heritage speakers of Spanish). Table 3.1 contains a summary of all these corpora, with indication of whether the corpus is a baseline corpus for the target language (i.e., Brazilian Portuguese), or a baseline for a source or potential transfer language (i.e., English and Spanish). I also discuss in this chapter the specific quantitative methods I use and how the data has been coded for these methods.

I use two quantitative methods of language analysis: 1) linear regression with word embeddings, which has been used in NLP studies and in quantitative social science research (Garg, Schiebinger, Jurafsky, & Zou, 2018) but not in acquisition studies, and 2) logistic regression, which has been extensively used in Language Variation and Change (LVC) (Tagliamonte, 2012) and in some L2 acquisition research (Geeslin & Guijarro-Fuentes, 2006; Woolsey, 2008).

For both of these methods I operationalize copula choice as *estar* preference for corpora in Portuguese and Spanish, with word embedding representing this preference as a continuous numeric variable and logistic regression modeling representing this preference as a binary variable (i.e., 0 for *ser* and 1 for *estar*). I expand more on this later in this chapter.

Since English has only one copula, i.e. *be* which maps to both *ser* and *estar* in Portuguese, the measure of interest is whether words in English that are similar to Portuguese words are more closely related to copula constructions than English words that are not similar to Portuguese words in form (Research Question #1). I will be using the term cognate in this dissertation to refer to words that are similar in orthographic form between two languages. For example, the word *gay* has the same orthographic form in both English and Portuguese. The word *cheap* in English is equivalent to *barato* in Portuguese, which has a very different orthographic form (i.e., there is no match between the sequence of orthographic letters in both languages). Equivalent forms for Portuguese and English were obtained using the R package `translateR` (Lucas & Tingley, 2014) through Google Translate API. I then used Levenshtein distance with the `adist()` function in R to calculate the difference in orthographic

form between two words.

For the L3 Portuguese data, the cognate (i.e., orthographic form similarity) measure is of English cognate vs. non-cognate words in Portuguese across different Spanish-English bilingual groups and L3 Portuguese level. I focus on adjectives, since adjectival predicates in Portuguese offer a choice between *ser* and *estar*. Word embeddings are used to compare distance range between 1) English copula *be* and Portuguese cognate and non-cognate adjectives in English across different English corpora (Research Question #1), and 2) L3 Portuguese copula *ser* and *estar* and English cognate and non-cognate adjectives in Portuguese across different Spanish-English bilingual groups and L3 Portuguese levels (Research Question #2).

Due to the need to analyze the data in an aggregated manner, the word embeddings provide an overall view of *estar* preference across the different corpora and predicate types. The logistic regression modeling allows for more details across different L1 backgrounds and L3 proficiency levels. The main goal of using these two methods is to triangulate results. The data is divided the following way, to answer each research question, based on my hypotheses:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?
 - Corpora: baseline corpora for Brazilian Portuguese, Spanish, and English
 - Hypotheses:
 - Spanish spoken in Arizona display different patterns of copula use from Spanish spoken in Spain and in Mexico (Bessett, 2015; Cortés-Torres, 2004; Geeslin & Guijarro-Fuentes, 2008; Salazar, 2007)
 - The patterns of copula use differ between Arizonian Spanish and Brazilian Portuguese (Geeslin & Guijarro-Fuentes, 2006; Moura, 2016; Sibaldo, 2011)
 - L1 English monolingual corpora display a preference for adjectives in copula *be* predicate position that are non-cognate with Portuguese
2. How do patterns of L3 Portuguese production by Spanish-English bilinguals compare

to the source languages (i.e., Spanish and/or English) versus the target language (i.e., Portuguese) at each of the three levels of L3 proficiency?

- Corpora: L3 learner Portuguese corpus divided into three Spanish-English bilingual groups and three proficiency levels
- Hypotheses:
 - If there is transfer from Spanish, copula use patterns in L3 Portuguese produced by Spanish-English bilinguals in their first semester of Portuguese is more similar to Arizonian Spanish patterns (Alonso & Rothman, 2016; Cabrelli Amaro, 2017; Carvalho & Bacelar Da Silva, 2006; Child, 2014; De Angelis, 2007; Forsyth, 2014; Rothman, 2014)
 - If there is no transfer from Spanish, L3 Portuguese learners will overuse *ser* (Garavito & Valenzuela, 2006; Ryan & Lafford, 1992; Schmitt & Miller, 2007; VanPatten, 1987, 2010)
 - If there is transfer from English, L3 Portuguese learners will show a preference for English cognate adjectives in copula predicate position (Ecke, 2015; Ecke & Hall, 2014; Hall et al., 2009)
 - At the last level of proficiency, L3 Portuguese patterns approximate those of Brazilian Portuguese baseline (De Angelis, 2007; Slabakova & Pilar García Mayo, 2017)
 - Acquisition differs across the three types of Spanish-English bilinguals because language status (i.e., L1 versus L2) is a factor language transfer from Spanish and English onto L3 Portuguese (Bardel & Falk, 2007; Bohnacker, 2006; Falk & Bardel, 2010; Håkansson et al., 2002; Schwartz & Sprouse, 1994, 1996)

Table 3.1: Summary of corpora used in this dissertation.

corpus	language	type	word.count
C-ORAL Brasil	L1 Portuguese	Baseline (target)	208,130
PRESEEA Spain	L1 Spanish	Baseline (transfer)	760,929
PRESEEA Mexico	L1 Spanish	Baseline (transfer)	597,916
CESA	L1 Spanish	Baseline (transfer)	498,711
The Cambridge Spoken	L1 English	Baseline (transfer)	1,192,527
BangorTalk Miami	L1 English	Baseline (transfer)	242,475
CORE	L1 English	Baseline (transfer)	2,631,582
MACAWS	L3 Portuguese	Learner Corpus	536,168

3.2 Baseline Corpora

3.2.1 Portuguese

For the Portuguese patterns of use by native speakers, I will be using the C-ORAL-BRASIL corpus (Raso & Mello, 2012), which consists of 139 spoken texts and over 21 hours of speech (208,130 words). This spoken corpus was designed to be representative of informal spoken Brazilian Portuguese, representing a variety of speech acts performed in everyday language. The data were recorded from spontaneous speech performances (e.g. conversation among friends, teachers’ meeting) in natural environments (Raso & Mello, 2012). The corpus is balanced in terms of context (divided into private and public) and “domains” (monologues, dialogues, conversations). Regarding varieties of Portuguese present in this corpus, 65% of the speakers are from Minas Gerais (a state in the north of the Southeastern region of Brazil), mainly from the metropolitan area of its capital, Belo Horizonte. Background information on each participant has also been collected, including age, sex, profession, etc. The language in the corpus was produced by a total of 511 participants, including 225 men and 281 women (no gender information is provided for 5 participants).

Table 3.2: Distribution of participants by age group and gender for C-ORAL-BRASIL (i.e., L1 monolingual Brazilian Portuguese)

age_group	female	male
Age Group A (18-25)	88	47
Age Group B (26-40)	65	73
Age Group C (41-60)	53	45
Age Group D (over 60)	7	15
Age Group M (minor)	6	6
x (unknown)	62	39

Table 3.3: Distribution of participants by education level and gender for C-ORAL-BRASIL (i.e., L1 monolingual Brazilian Portuguese)

education	female	male
Primary school not completed	26	32
Up to graduated that does not or did not use the graduation for her/his job	103	59
Graduated who uses the graduation for her/his job or higher	90	93
Unknown	62	40

Table 3.4: Word count distribution across genres in the web corpus of Brazilian Portuguese.

genre	description	word_count
advice	beauty and home decoration advice	20151
daily blog	narratives on daily life events	11567
food blog	food product and restaurant reviews, and recipes	12975
media blog	book, tv show, and movie reviews	14517
opinion blog	blog posts with opinions on politics and current events	12178
short stories	short narratives	18475
travel blog	narratives on travel experiences	144315

3.2.2 Spanish

For Spanish produced by native speakers, I will use two main corpora: the Corpus del Español en el Sur de Arizona (CESA - Corpus of Spanish in Southern Arizona) (Carvalho, 2012), the PRESEEA Corpus (Proyecto Para El Estudio Sociolingüístico del Español de España y de América) (PRESEEA, 2014).

CESA consists of 76 transcribed sociolinguistic interviews carried out by students enrolled in undergraduate and graduate sociolinguistics courses. Background information on each participant is also collected in two forms: 1) Bilingual Language Profile of the Participant (BLP), which contains information about their English and Spanish language learning and use; and 2) Demographic Information of the Participant (DI), which contains information such as age, sex, education, etc. CESA informants are between the ages of 18 and 68 (mean = 27.8, SD = 11.2), and include 48 women and 28 men. See Table 3.5 for the distribution of age across sex in CESA. Regarding education, CESA is more heavily represented by college-educated informants with 36% of participants holding either a bachelor's or associate's degree, and 50% were students in university, at the time of the interview.

Table 3.5: Distribution of participants by age group and gender for CESA (i.e., bilingual Spanish in Southern Arizona).

age_group	Female	Male
Age Group A (18-25)	29	20
Age Group B (26-40)	8	5
Age Group C (41-60)	8	2
Age Group D (over 60)	0	1
x (unknown)	3	0

Table 3.6: Distribution of participants by education level and gender for CESA (i.e., bilingual Spanish in Southern Arizona).

Education Level	Female	Male
Associate's Degree	2	2
Bachelors Degree	11	12
High School	6	3
Master's Degree	1	0
sixth grade	1	0
some college	27	11

For any sociolinguistic interview used in this dissertation, only the language produced by the informants (not the interviewers) is analyzed. For CESA, a total of 498,711 words were produced by informants.

For the other Spanish baseline corpora, I used two corpora from the PRESEEA project, whose main goal was to coordinate sociolinguistic investigations of Spanish spoken in the

Americas and in the Iberian peninsula through shared protocols for selecting speakers and collecting data. Each semi-structured interview lasts around 45 minutes, and the general interview structure is as follows: greetings, the weather, place where they live, family and friendships, habitual activities and future plans, near-death or dangerous past experiences, important life anecdotes, wishes of economic improvement, closing.

Examples of prompts for each one of these modules are shared among interviewers. The goal of this interview protocol is to allow participants to relax from a more formal to a more informal and natural speech (Fernández, 1996). The topics also allow opportunities for continuous participant speech that encompass a broad range of language structures (e.g., present, past, future, daily activities, imagined situations).

In this dissertation, two subcorpora of the PRESEEA Corpus (1993-), Mexico and Spanish, are the baseline for Spanish produced by L1 monolinguals. The Spanish corpus is comprised of 97 interviews (760,929 words), while the Mexico corpus contains 67 interviews (597,916 words). These two locations were chosen to establish contrasts between copula use in continental Spanish, versus the closest L1 monolingual variety (i.e., Mexico) to the type of Spanish spoken in Arizona. Participants for this subcorpora were between the ages of 20 and 89, and are comprised of 58 women and 56 men (see Table 3.7 for a distribution of participants' ages across gender). In terms of level of education, PRESEEA is more evenly represented across three levels: low, medium, and low (Table 3.8).

Table 3.7: Distribution of participants by age group and gender for PRESEEA subcorpora (i.e., Mexico and Spain).

age_group	Female (Mexico)	Female (Spain)	Male (Mexico)	Male (Spain)
Age Group A (18-25)	5	6	7	10
Age Group B (26-40)	11	15	7	11
Age Group C (41-60)	10	16	13	17
Age Group D (over 60)	6	10	8	12

Table 3.8: Distribution of participants by education level and gender for both PRESEEA subcorpora (i.e., Mexico and Spain).

education	Female (Mexico)	Female (Spain)	Male (Mexico)	Male (Spain)
Up to 5 years of formal education	11	16	11	16
High School	10	15	12	17
University or associate’s degree	11	16	12	17

Table 3.9: Summary of the oral Spanish baseline corpora used.

corpus	type	interviews	word_count
PRESEEA Spain	Monlingual Spanish	97	760929
PRESEEA Mexico	Monlingual Spanish	67	597916
CESA	Bilingual Spanish	76	498711

3.2.3 English

English produced by native speakers is represented by three corpora as well: The Cambridge International (Cambridge, 2004), and the Corpus of Online Registers of English (CORE) (Biber, Egbert, & Davies, 2015), and the BangorTalk Miami (Deuchar, 2011) corpora.

The BangorTalk Miami corpus consists of conversations by Spanish-speakers in Florida, all of whom are bilingual in English (Deuchar, 2011). The corpus is comprised of 56 informal conversation between two or more speakers, for a total of 84 speakers living in Miami, Florida (35 hours of speech, 242,475 transcribed words). Spanish utterances represent 12% of the corpus (28,689 words), but only the English tokens will be used for this corpus. Information on each participant such as age, gender, places lives, is also available.

The Cambridge Spoken corpus is comprised of conversations by 733 participants, from a variety of US states, representing a variety of Americans from different age groups, and education backgrounds. Table 3.10 shows the distribution of participants' age across gender, and Table 3.11 shows the distribution of participants' education level across gender. This corpus seems to be overrepresented by Age Group B (26-40) participants with a tertiary level of education (i.e., Level 3).

The CORE corpus (Biber et al., 2015) consists of eight registers in spoken and written English produced by native speakers, for a total of 50 million words. I selected sub-register categories that represent the types of written texts found in the learner corpus used in this dissertation (see description in the next section), such as Travel Blog, Short Story, and Advice, which totals 2,631,582 words. This corpus has been previously tagged with the Biber tagger (Biber, 2006) and tag-checked.

Table 3.10: Distribution of participants by age group and gender for the Cambridge Spoken corpus.

age_group	Female	Male	Unknown
Age Group A (18-25)	24	36	0
Age Group B (26-40)	125	59	0
Age Group C (41-60)	87	29	0
Age Group D (over 60)	12	7	0
Child (under 18)	10	22	0
x (unknown)	45	59	218

Table 3.11: Distribution of participants by education group and gender for the Cambridge Spoken corpus .

education	Female	Male	Unknown
Level 1	13	13	0
Level 2	36	48	0
Level 3	136	63	0
Level 4	71	26	0
Level Unknown	47	62	218

3.3 Learner Corpus

The Portuguese learner corpus used in this dissertation is comprised of 2,075 assignments (536,168 tokens) from 15-week face-to-face Portuguese language courses at University of Arizona. This corpus includes assignments only from students who have signed informed consent, and data was collected from five semesters (Spring 2017 to Spring 2019). Participants also fill out a survey with language background information, so they are labeled as belonging to one of the three Spanish-English bilingual groups.

The corpus contains language produced by 255 students who have agreed to share their Portuguese language course assignments for this project. 30% of these students have signed the informed consent form for multiple semesters (n=77). Self-reported language background is collected through a survey, where participants list their native and additional languages, and check a statement that best describes their experience with Spanish. At times, a learner will say they speak Spanish as an L2 but they also check the statement that says “I was born in a Spanish-speaking country and lived there until at least 5 years old” or “I was exposed to Spanish as a child in my household in the US” – these participants are labeled as L1 Spanish speakers in the corpus. There are also learners who say English is their L2, but declare that they grew up in the US – these participants are labeled as L1 English speakers.

For this dissertation, participants are divided into three bilingual groups: 115 (45%) are heritage speakers of Spanish (L1 English L1 Spanish), 79 (31%) are L1 English L2 Spanish speakers, and 38 (15%) are L1 Spanish L2 Speakers. The rest of the corpus participants (n=23, 9%) either were not Spanish-English bilinguals (n=8, 3%) or did not provide their language background (n=15, 6%), and are thus excluded from this dissertation.

Assignments have been collected from the following Portuguese language courses:

Table 3.12: L3 Portuguese learner corpus summary of document count, total word count, and mean document length in words per course.

course	document_count	total_words	mean_document_length
PORT 305	1009	140271	139.02
PORT 325	555	150880	271.86
PORT 350	40	27273	681.82
PORT 425	493	219230	444.69

Spanish is a common language at University of Arizona, which has recently earned the designation of Hispanic-Serving Institution from the U.S. Department of Education. The percentage of Hispanic students enrolled at University of Arizona has increased in the past decade (Figure 3.1), with more than a quarter of current undergraduate students at this institution identifying as Hispanic.

At University of Arizona, all students in B.A. degree programs are required to achieve a fourth-semester skill level in a foreign language of their choosing. For all other undergraduate students (i.e., BS degree seeking students), a second-semester foreign language proficiency level is required instead. For Spanish-English bilinguals, they need to achieve that level of proficiency in a language other than Spanish. In addition, undergraduate students who major in Spanish are required to take two Portuguese language courses, which can also satisfy their foreign language requirement. Minors in Portuguese need to take a total of 18 credit units.

Due to these foreign language requirements, the Portuguese language program at University of Arizona is one of the largest in the country (Looney & Lusin, 2019).

Every semester, the Portuguese language program offers the following undergraduate courses.

- PORT 305 – First semester Portuguese for Spanish speakers
- PORT 325 – Intermediate Portuguese conversation (second semester)
- PORT 425 – Advanced Portuguese composition (third semester+)
- PORT 350 – Introduction to genres and literary analysis (third semester+)

All sections of the courses above are taught by graduate students enrolled in a PhD program in either the Department of Spanish and Portuguese, or in Second Language Acquisition and Teaching, an interdisciplinary doctoral program. Regular first and second semester Portuguese courses (i.e., PORT 101 and 102, for non-Spanish speakers) are occasionally offered, depending on enrollment numbers. In addition to these language courses, Portuguese courses in linguistics (e.g., Portuguese Phonology), literature (e.g., Topics in Luso-Brazilian Literature), and anthropology (e.g., Brazilian Identity: Class, Race, and Citizenship) are taught by faculty from the Department of Spanish and Portuguese and the Center for Latin American Studies.

As mentioned, students in the Portuguese language program fill out a survey where they self-report their language background.

Regarding the students who consented having their assignments collected for the corpus, 88% are between the ages of 17 and 20 (Table 3.13). Gender information is not collected. When it comes to year in school, the bulk of participants across semesters are sophomores and juniors (Table 3.14).

Table 3.13: Age distribution of L3 Portuguese corpus participants.

age_group	n
Age Group A (18-24)	215

age_group	n
Age Group B (25-40)	21
Age Group C (over 40)	3

Table 3.14: Year in school distribution of L3 Portuguese corpus participants.

year	n
Freshman	19
Sophomore	76
Junior	107
Senior	107
Graduate	13

The assignments for the Portuguese language program are submitted through a course management system, and the intended audience is the instructor only. For this dissertation, only data from PORT 305, PORT 325, and PORT 425 are analyzed, since these three courses are taken in sequence. Each one of these courses is used as proxy for proficiency level, with first semester Portuguese (PORT 305) representing the lowest level of proficiency, and Portuguese composition (PORT 425) the highest (or third) level of proficiency.

- PORT 305: Biweekly written assignments, topic based (e.g., describe your family).
- PORT 325: Weekly written assignments, online discussions (e.g., choose a work of art that you like, describe it, and explain what you like about it)
- PORT 425: Genre-based written assignments (e.g., gastronomic memory, trip report, news article)

The texts used for this dissertation are written assignments only. These were untimed, and written at home (i.e., not during class time). For some assignments, not all, students submitted a draft before submitting the final version.

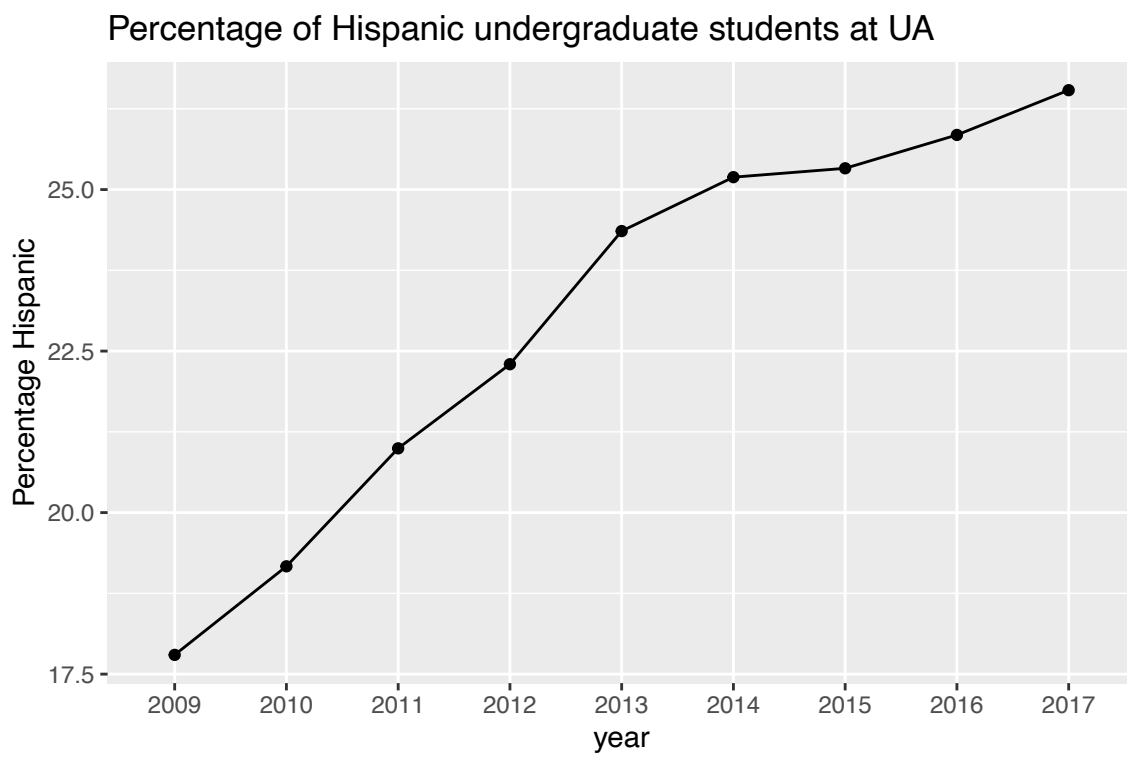


Figure 3.1: Percentage of undergraduate Hispanic students enrolled at University of Arizona across the years (2009-2017).

3.4 Analysis

3.4.1 Data Extraction

For all baseline spoken corpora, I used the orthographic transcriptions provided by the corpora creators. Copula verbs are identified using different labeling tools for each language or corpus. The creators of the C-ORAL-BRASIL corpus provide transcriptions for all spoken texts, which are in CHAT format (MacWhinney, 2014). The transcriptions come tagged with PALAVRAS parser (Bick, 2000), a rule-based tagger for Portuguese. A total of 9,102 instances of copula use were extracted from this corpus. For the other two Portuguese corpora, I used the fine-grained part of speech labels produced by the RNNTagger (Schmid, 2019). A total of 10,431 instances of copula were extracted from the L3 Portuguese corpus, i.e., MACAWS (Staples, Novikov, Picoral, & Sommer-Farias, 2019).

For the CORE corpus, I used the labeled files, which have been tagged with the Biber tagger (Biber, 2006). All other corpora were tagged with the Stanford dependency parser (Chen & Manning, 2014), with the English (Schuster & Manning, 2016) and Spanish (Gauthier, 2016) configuration options for the respective languages. A total of 5,860 instances of copula use were identified from CESA, 7,355 from PRESEEA Spain and 4,745 from PRESEEA Mexico. For the English corpora, 6,309 instances of copula use were extracted from the BangorTalk Miami corpus, and 73,922 from CORE.

Once the corpora constructions were extracted, further labeling of each instance proceeded according to the quantitative method to be used. These processes included both automatic labeling with Python scripts written by the author, or by hand. Further details of this is discussed under each quantitative method section that follows.

Table 3.15: Distribution of copula instances across corpora.

Corpus	Word Count	Annotation Tool Used	Copula Instances
C-ORAL Brasil	208,130	PALAVRAS parser	9,102
PRESEEA Spain	760,929	Stanford dependency parser	7,355
PRESEEA Mexico	597,916	Stanford dependency parser	4,745
CESA	498,711	Stanford dependency parser	5,860
The Cambridge Spoken	1,192,527	Stanford dependency parser	33,615
BangorTalk Miami	242,475	Stanford dependency parser	7,492
CORE	2,631,582	Biber Tagger	73,922
MACAWS	536,168	RNNTagger	10,431

3.4.2 Quantitative Analysis

Individual words are coded into different categories following their part of speech tag and according to previous literature. Adjectives in English and Portuguese were divided in two classes: cognate and non-cognate with Portuguese and English respectively. For adjectives in Spanish and Portuguese, the following categories were used:

- age (e.g., velho/viejo/old) (Bessett, 2015; Cortés-Torres, 2004; Salazar, 2007; Silva-Corvalán, 1986)
- color (e.g., verde/green) (Geeslin & Guijarro-Fuentes, 2008)
- physical appearance (e.g., feio/feo/ugly) (Bessett, 2015; Cortés-Torres, 2004; Geeslin & Guijarro-Fuentes, 2008; Salazar, 2007)
- (other) description (e.g., perigoso/peligroso/dangerous) (Bessett, 2015; Geeslin & Guijarro-Fuentes, 2008; Salazar, 2007; Silva-Corvalán, 1986)
- emotional/mental state (e.g., alegre/happy) (Bessett, 2015; Ramirez-Gelpi, 1997)
- evaluation (e.g., bom/bueno/good) (Bessett, 2015; Cortés-Torres, 2004; Geeslin & Guijarro-Fuentes, 2008; Salazar, 2007; Silva-Corvalán, 1986)

- order (e.g., primeiro/primerero/first)
- sensory (e.g., doce/dulce/sweet) (Geeslin & Guijarro-Fuentes, 2008; Salazar, 2007; Silva-Corvalán, 1986)
- shape (e.g., quadrado/cuadrado/square)
- size (e.g., pequeno/pequeño/small) (Bessett, 2015; Cortés-Torres, 2004; Geeslin & Guijarro-Fuentes, 2008; Salazar, 2007; Silva-Corvalán, 1986)
- social/ethnic class or classifier (e.g., católico/catholic) (Bessett, 2015; Salazar, 2007; Silva-Corvalán, 1986)
- verbal (e.g., cansado/tired) (Garavito & Valenzuela, 2006)
- miscellaneous (Bessett, 2015)

As mentioned, the two quantitative methods of language analysis used in this dissertation use as their dependent variable a measure of either *estar* or *cognate* preference, operationalized differently for each method. For *cognate* preference, first words were translated using Google’s cloud translation API (English to Portuguese, and Portuguese to English and Spanish). Then, approximate string distances between original word and translations were calculated in R using the `adist()` function, which uses a generalized Levenshtein (edit) distance, giving the minimal possibly weighted number of insertions, deletions and substitutions needed to transform one string into another.

The quantitative methods used are applied to subsets of the data, filtered by word category (e.g., size adjectives, *em* prepositions). For example, only *copula + size adjectives* constructions are selected, then both linear regression with word embeddings and logistic regression are run on the data. The two methods of analysis employed (i.e. word embeddings and logistic regression) provide different but complementary perspectives on copula use. In the sections below, each method is explained in more detail.

3.4.2.1 Word Embeddings

This unsupervised Machine Learning technique creates a vector (i.e., a list of 100 numbers) to represent each word in a corpus. These vectors are then used to understand word-to-word

similarities based on context (Zou, Socher, Cer, & Manning, 2013). That is done through calculation of distances between words, with similar words being grouped together, i.e., their vectors are close in distance compared to other non-related words in the corpus (Goldberg & Levy, 2014; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Zou et al., 2013). These vector distance calculations have been shown to reveal attested linguistic regularities and patterns. For example, the calculation $vector("Madrid") - vector("Spain") + vector("France")$ results in a vector that is closer to $vector("Paris")$ than to any other word vector in the corpus (Mikolov et al., 2013).

There are two main architectures for word embeddings: 1) Continuous Bag of Words (CBOW) and 2) Skip-gram; the difference being that CBOW does not take into account word order, and Skip-gram does (Levy & Goldberg, 2014; Mikolov et al., 2013). For this dissertation, I use the skip-gram architecture since word order is important. The specific algorithm used in this dissertation is Word2vec (Goldberg & Levy, 2014; Mikolov et al., 2013), implemented in Python using the Gensim library (Řehůřek & Sojka, 2010). Besides architecture, the size of the vectors, the minimum frequency rate, and the size of the training window are parameters that affect the performance of this algorithm (Mikolov et al., 2013). That means I have trained the model on my data several times, changing these parameters systematically, in order to decide on the best parameters to be used in the final model. A window size too large took into account phrasal units that were too distant from the copula verb, which in turn caused attributive adjectives (e.g., *esse aí é um cara bonito*, *that one is a handsome guy*) to be conflated with predicative adjectives (e.g., *esse cara é bonito*, *this guy is handsome*). Thus, I decided on a small window of size 2. Due to the corpora being under 1 million words in size, I set the minimum frequency of 5 occurrences per token. I decided on 100 numbers for the size of the vectors, because of the slowing down during Euclidean distance calculation with larger vectors. Once the hyperparameters were established, the algorithm was run using the same parameters across corpora, outputting a set of embeddings for each corpus.

Each copula verb was tagged with a *cop* label, based on the copula extraction process described earlier, to differentiate them from other types of verbs. In addition, each token was also combined with its part of speech tag (e.g. *bonito_ADJ*) also for differentiation

between form and function. Based on the resulting embeddings, which comprised a list of 100 numbers for each *token_tag* in the corpus, Euclidean distances between target lexical items (e.g., adjectives and adverbs) and all forms of copula *ser* and *estar* were calculated, which were then used to measure *estar* preference (i.e., distance to *ser* minus distance to *estar*) for each word. With Euclidean distances, the larger the number, the further apart two embeddings are. My main goal here is to identify which words are closest to each of the copula verbs in each of the baseline corpora in all three languages (Research Question #1). I am also using word embeddings with the learner corpus data to show changes in word associations (e.g., verb choice + adjective) across different proficiency levels (Research Question #2).

After distances between copula forms and different words in the corpus were calculated, preference for *estar* across corpora was established based on the difference between word embeddings for *ser* and *estar* and the target lexical items in a predicate category (e.g., adjectives of size). For example, the Euclidean distance between *é* (*ser* copula verb) and *bonito* (adjective) is calculated (distance A), then the distance between *está* (*estar* copula verb) and *bonito* (the same adjective) is calculated as well (distance B) and then subtracted from the *ser* copula distance previously calculated (distance A - distance B). Positive numbers indicate a preference for *estar*, while negative numbers, a preference for *ser*. Cognate preference across corpora was established based on the word embedding distances between all forms of copula verbs and adjectives divided by cognate status (i.e. cognate and non-cognate).

To determine significant associations between *estar* preference and each group (e.g., baseline corpora, proficiency level), linear regression models for each type of predicate (e.g., size adjectival, *em* prepositional) were run in R using the lme4 package (Bates, Mächler, Bolker, & Walker, 2014) with *estar* preference as the dependent variable, corpus as independent variable, and word lemma as a fixed effect. Categorical factors (e.g., L3 level, L1 background) are always coded as numbers in regression modeling, with the default in R outputting the estimate of one factor level (e.g., Level 2, Level 3) in relation to another factor level (e.g., Level 1). That is called treatment coding, because one factor level is the control group, and the additional factor levels are the different treatments in an experimental setting (Winter, 2019).

Since corpus data is observational data, there is no control group to be used as the reference level for the regression. As such, the contrast matrix for the regression modeling is set as its identity matrix. The linear regression models are thus without contrasts (i.e., deviation from zero), with each factor presenting its own estimate mean (as opposed to having one factor level as the baseline) (Bolker, 2018; Fox & Monette, 2002). That is possible because all *estar* preferences (i.e., the dependent variable) are centered on zero. In this setting, significant p values (i.e., $p < .05$) indicate whether that group mean is significantly different from zero (i.e., whether there is a preference for *estar* or *ser*). Difference between groups is established through confidence intervals. Regarding assumptions for these, Figure 3.2 shows the overall distribution of *estar* preference in the data, which shows that *estar* preference is mostly symmetrical, uni-modal, and normally distributed across corpora.

3.4.2.2 Logistic Regression

Estar preference can also be approached as a binary choice (i.e., *ser* or *estar*) in logistic regression modeling. As mentioned, all texts were tagged for part of speech and syntax information such as tense-mood-aspect, person, gender and number. Each instance of copula use was then extracted from the corpora and labeled as being either an instance of *ser* or *estar* use. Then, based on the parts of speech patterns of the words that come after the copula verb, an R script was written by the author to label each construction for type of predicate, e.g., *copula + adverb + adjective* constructions are labeled as adjectival predicate. These are then hand-checked to eliminate any errors in this automated labeling process. Due to this labor-intensive hand-checking process, results from only one of the Spanish corpora is reported for logistic regression modeling. The purpose of the regression modeling is to shed light on patterns of copula use for each baseline group (Research Question #1), and learner group at each proficiency level (Research Question #2).

The goal of this type of analysis is to understand the preference for *estar* or for cognate adjectives according to internal (i.e., linguistic) and external (i.e., extra-linguistic or social) factors (Tagliamonte, 2012). In this approach, participant is included as a random effect (Gries & Deshors, 2014; Norris, 2015) in the logistic regression modeling. Different independent

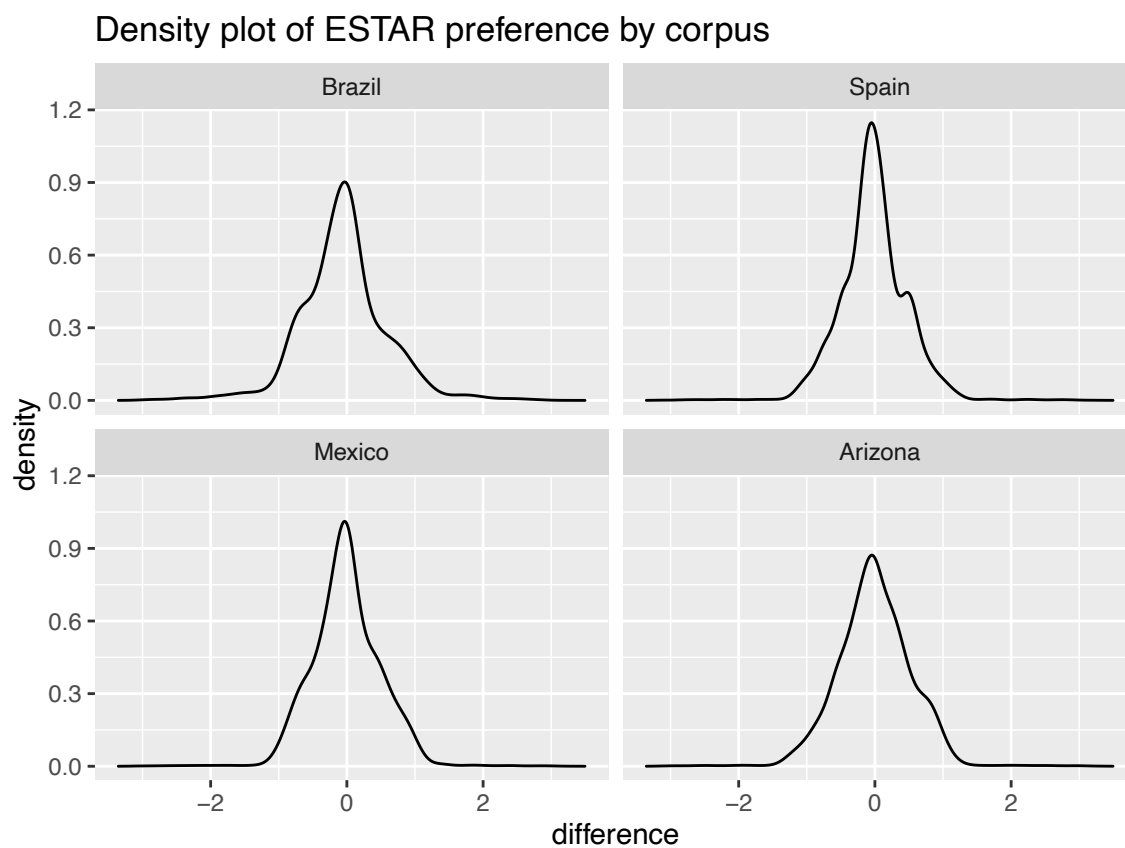


Figure 3.2: Density distribution of ESTAR Preference with all lexical items across languages (N=474,572). Charts are centered on zero. Positive numbers (right from zero) represent a preference for ESTAR. Negative numbers (left from zero) represent a preference for SER. Distributions are mostly symmetrical, uni-modal, and normally distributed for all corpora

variables (or factor groups) are added to the models, including tense-mood-aspect (TMA) of the copula verb, subject type (e.g., pronoun vs. noun phrase), and predicate type (e.g., adjective, noun phrase, clause). Extra linguistic factors include L1 (i.e., L1 English, Heritage Spanish, L1 Spanish) and L3 Portuguese level (i.e., Level 1, Level 2, and Level 3). The main goal here is to understand what factors (e.g., types of predicates) favor *estar* use or *cognate* use. The logistic regression outputs log odd estimates that are then transformed in probabilities. These estimates reflect the probability of *estar* or *cognate* use for each predicting factor included in the regression model. Again here we are dealing with corpus data, which is observational data. As mentioned in the linear regression section above, the default in R for factor level testing is treatment contrast coding (Winter, 2019), which is not appropriate here. The default for logistic regression from a variationist approach is sum coding contrast, which uses a group mean as the contrast for all factor levels (Tagliamonte, 2012). The sum coding approach assumes that the different groups have a mean in common because they are all part of the same speech community, which is not the case for the data in this dissertation, which spans across different languages and language varieties. As such, the contrast matrix for the regression is set as its identity matrix, with significant p values (i.e., $p < .05$) meaning that estimate probabilities are significantly different than pure chance (i.e., different from a 50% probability of *estar* use).

3.5 Conclusion

The methods employed in this dissertation include the use of an L3 Portuguese corpus distributed across three Spanish-English bilingual groups across three L3 Portuguese levels, and baseline corpora in Portuguese, Spanish and English. The quantitative methods chosen are meant to provide a complementary view of copula use across the different corpora. The goal is to shed light on *estar* preference with different types of copula predicate across the different Spanish and Portuguese corpora, and on *cognate* preference across the different English and Portuguese corpora.

Estar preference is operationalized in two ways. The first uses word embeddings, with *estar*

preferences defined as the the distances between copula *ser* embeddings and the target (e.g., size adjectives) embeddings minus the distances between copula *estar* and the same target embeddings. The second method is logistic regression, with the use of either *ser* or *estar* in a copula construction used as the binary dependent variable. Each of the two research questions is answered by applying both methods across different corpora. *Cognate* preference is established through word embedding distances with cognate vs. non-cognate words, and as a binary variable (e.g., cognate vs. non-cognate) for logistic regression modeling.

To answer the first research question, I analyze the baseline corpora. For the second research question, the L3 Portuguese Learner corpus is analyzed with differences and similarities in results being drawn across the three Spanish-Bilingual groups and the tree L3 Portuguese levels.

CHAPTER 4

RESULTS 1 – EMBEDDINGS FOR SPANISH CORPORA

4.1 Introduction

This first results chapters addresses partially the following question:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?

This question is answered in three parts. In this chapter I discuss word embedding results that show that copula use in Spanish spoken in Southern Arizona differs from Spanish spoken in Spain and in Mexico. In the next chapter, I discuss *estar* embedding preference differences between Arizonian Spanish and Brazilian Portuguese, and Portuguese cognate adjective preference with *be* copula embeddings in American English (another potential source of language transfer).

As previously discussed, preference for *estar* embeddings across corpora is established based on the difference between word embeddings for *ser* and *estar* and target lexical item embeddings in a predicate category (e.g., adjectives of size). Positive numbers indicate a preference for *estar*, while negative numbers, a preference for *ser*.

When discussing significant differences between corpora, I present statistical inference results for linear regression models for each type of predicate (e.g., size adjective, intensifiers), which were run with *estar* embedding preference as the dependent variable, corpus as independent variable, and word as a fixed effect (i.e., random intercept). The linear regression models are reported without contrasts (i.e., deviation from zero), with each level (i.e., corpus) presenting its own estimate mean (as opposed to having one factor level as the baseline, a.k.a. treatment contrasts) (Bolker, 2018; Fox & Monette, 2002). That is possible because all *estar* embedding

preferences (i.e., the dependent variable) are centered on zero. The p values presented in tables indicate whether that group mean is significantly different from zero (i.e., whether there is a preference for *estar* or *ser*). Difference between groups is established through confidence intervals. Results are divided by types of predicates.

4.2 Spanish in Arizona, Mexico, and Spain

In this section, I discuss the differences in copula embeddings between Spanish in Southern Arizona and Spanish spoken in Mexico and Spain. The L3 Portuguese learners and Spanish-English bilinguals who are participants in this research are from Southern Arizona, so Spanish spoken in Southern Arizona is represented by the CESA corpus. Spanish from Mexico, the closest variety to the Spanish spoken in Southern Arizona, and Spanish from Spain are both used to highlight particular patterns of copula use in these corpora.

I focus on lexical categories that have been shown to differ across Spanish varieties, mainly from studies on the extension of *estar* in Spanish spoken in the United States (Bessett, 2015; Cortés-Torres, 2004; Salazar, 2007; Silva-Corvalán, 1986). As discussed in the literature review, while *estar* extension is not universal for all bilingual groups, a number of linguistic predictors of *estar* are common across groups (Geeslin & Guijarro-Fuentes, 2008). Adjectives of size have been shown to favor *estar* in Spanish spoken in Los Angeles (Cortés-Torres, 2004; Silva-Corvalán, 1986). Adjectives of age also favor *estar* in Spanish in New Mexico (Salazar, 2007) and in Los Angeles (Cortés-Torres, 2004; Silva-Corvalán, 1986). Physical appearance and evaluation adjectives favor *estar* in Spanish in Los Angeles (Silva-Corvalán, 1986). Regarding adverbs modifying adjectives in predicate position, the presence of an intensifier have been shown to favor *estar* in New Mexico (Salazar, 2007). More relevant to this dissertation, Bessett (2015) found that adjectives of age, appearance, and size favor *estar* in Spanish spoken in Southern Arizona.

4.2.1 ESTAR Preference with Adjectives

4.2.1.1 Age Adjectives

Adjectives of age found in the corpora include words such as *joven* (*young*), and *viejo* (*old*). See Table 4.1 for examples of age adjectives found in the Spanish corpora and their corresponding translation to English.

Table 4.1: Adjectives of age in the Spanish corpora.

word	translation
<i>joven</i>	young
<i>mayor</i>	older
<i>mayores</i>	older
<i>nuevo</i>	young/new
<i>viejo</i>	old

Table 4.2 shows linear regression results for *estar* embedding preference with age adjectives as the dependent variable, and corpus as a predictor. Estimates above zero mean a preference for *estar* embeddings, below zero, a preference for *ser* embeddings. Confidence intervals are used to establish significant differences between corpora. As can be seen, the Arizona corpus displays a preference for *ser* embeddings with adjectives of age (see Figure 4.1 for visual representation of results), being significantly different from the Mexico corpus only. The Spain corpus does not display a preference for either *ser* or *estar* embeddings with adjectives of age ($p > .05$ and 95% CIs [-0.03, 0.02]). The preference for *estar* embeddings with adjectives of age for the Mexico corpus (95% CIs [0.00, 0.06], $p < .05$) confirms the presence of *estar* extension in this corpus. The preference for *ser* embeddings present in the Arizona corpus (95% CIs [-0.06, -0.01], $p < .05$) might be indicative of a more conservative use of copula in this corpus.

Table 4.2: Linear regression results with 95% confidence intervals for adjectives of age across Spanish baseline oral corpora.

corpus	Estimate	CI	Pr(> t)
Arizona	-0.03	[-0.06, -0.01]	0.03 *
Mexico	0.03	[0.00, 0.06]	0.04 *
Spain	-0.01	[-0.03, 0.02]	0.69

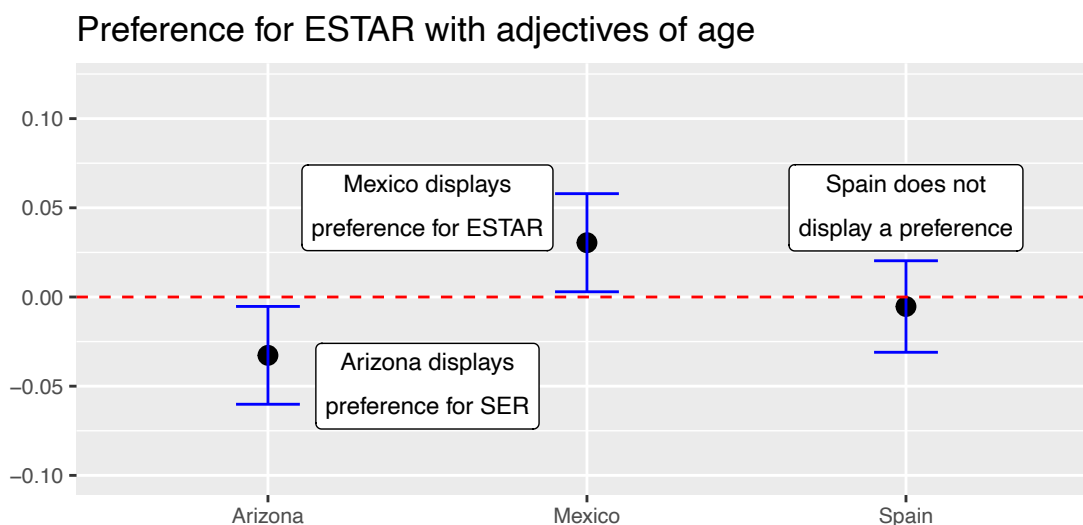


Figure 4.1: ESTAR Preference with adjectives of age across Spanish baseline oral corpora. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

The examples below illustrate the use of *ser* and *estar* with age adjectives in the Mexico corpus (1-2), the Spain corpus (3-4), and the Arizona corpus (5-6). As demonstrated below, the Mexico corpus displays uses of *estar* with words like *viejo/old* (1) and *mayor/older* (2) while also presenting *ser* with words like *nueva/young* (2). There are some instances of *estar* with age adjectives in the Spain corpus (4) and some where the speaker uses *ser* instead (3). Arizona shows a preference for *ser* with adjectives of age (5-6).

(1) ya **está viejo** ya tiene que emigrar para que llegue gente joven (Mexico DF - H21 090)

(you) are already old (you) already have to emigrate for young people to arrive

(2) síntomas del parto pues ya **está mayor** y **es nueva** pues (Mexico Mexicali - H21 023)

childbirth symptoms because (she) is already older and (she) is young

(3) yo he conocido también a otra persona con Parkinson que no **es mayor** que **es joven** (Spain Granada M31 053)

I have also met another person with Parkinson that is not older that is young

(4) yo a mis niños se lo digo que cuando ellos **estén mayores** que tienen que salir a tomarse unas cervezas (Spain Granada H21 043)

I to my kids say that when they are older that they have to go out and drink some beer

(5) son legalmente ciudadanos y todo, y **eran bien viejos** (CESA 072)

(they) are legal citizens and all that, and they were really old

(6) porque yo **era joven** cuando visitó (CESA 050)

because I was young when (he) visited

4.2.1.2 Size Adjectives

Adjectives of size found in the Spanish corpora include words such as grande (*big*), and pequeño (*small*). See Table 4.3 for examples of these size adjectives and their respective translation to English.

Table 4.3: Adjectives of size in the Spanish corpora.

word	translation
chica	little
chico	little
chiquita	little
chiquito	little
gran	big
grande	big
grandes	big
pequeña	small
pequeño	small
pequeños	small

Table 4.4 provides results for the linear regression with *estar* embedding preference as the dependent variable and corpus as the predictor. As shown, the Arizona corpus displays a preference for *ser* embeddings with adjectives of size, overlapping with Spain being significantly different from the Mexico corpus. Speakers in the Spain corpus also display a preference for *ser* embeddings with adjectives of size (see Figure 4.2 for a visual representation of these results). While these results parallel results for age adjectives discussed earlier, with Mexico being the only corpus displaying some *estar* extension with size adjectives (95% CIs [-0.01, 0.03]), they are not as strong since none of the results for all three corpora are significantly different from zero ($p < .05$).

Table 4.4: Linear regression results with 95% confidence intervals for adjectives of size across Spanish baseline oral corpora.

corpus	Estimate	CI	Pr(> t)
Arizona	-0.02	[-0.03, 0.00]	0.12
Mexico	0.01	[-0.01, 0.03]	0.22
Spain	-0.01	[-0.03, 0.01]	0.20

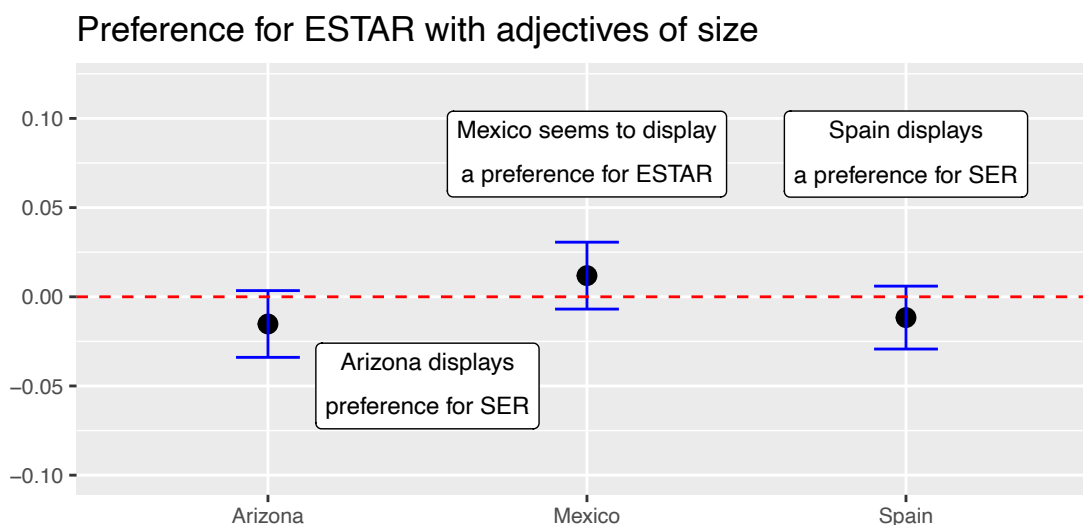


Figure 4.2: ESTAR Preference with adjectives of size across Spanish baseline oral corpora. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

The excerpts below show examples of *ser* and *estar* use with adjectives of size for the Mexico corpus (7-9), the Spain corpus (10-12), and the Arizona corpus (13-15). In the Mexico corpus, *estar* is used with words like *grande/big* and *chico/little* (7-8). The copula *ser* is also used with these adjectives (9) in the Mexico corpus. Both copula are used with adjectives of size in the Arizona corpus as well; see (13) for an example of a super token (Tagliamonte, 2012), where the speaker uses first *ser* with *grande/big* and then *estar*. The instances of *estar* use with adjective of size in the Spain corpus seem to occur mainly when people are talking about

children (12).

- (7) toda esa gente se drogaba cuando **yo estaba** chico (Mexico Guadalajara H13 014)

all those people got high when I was little

- (8) el recinto ferial de Puebla **está bien grande** ahí (Mexico Mexicali H32 018)

the fairground of Puebla is very large large there

- (9) le digo que el Distrito no **era grande** (Mexico DF H31 102)

I say that the District was not big

- (10) además **es muy pequeñita**, Mallorca sí **es grande** (Spain Malaga M22 704)

also it is very small, Mallorca is large

- (11) no el piso aún **es grande** (Spain Santiago de Compostela M12 020)

no the floor is still big

- (12) ahora **está más grande** con el pelo corto (Spain Madrid M11 004)

now she is bigger with the short hair

- (13) tenía como tres meses pero **estaba bien grandes** (CESA 063)

they were like three month old but they were very big

- (14) la comunidad **es grande** pero no **está muy grande** como otras comunidades en Estados Unidos (CESA 033)

the community is big pero it is not too big like other communities in the United States

- (15) me gustaba mucho cuando **era pequeña** (CESA 060)

I liked it very much when I was little

4.2.1.3 Evaluation Adjectives

Adjectives of evaluation found in the corpora include words such as *buena* (*good*), and *malo* (*bad*). See Table 4.5 for examples of evaluation adjectives found in the Spanish corpora and their respective translation to English.

Table 4.5: Adjectives of evaluation in the Spanish corpora.

word	translation
buen	good
buena	good
buenas	good
bueno	good
buenos	good
difícil	difficult
fácil	easy
horrible	horrible
mal	bad
mala	bad
malo	bad

Table 4.6 contains results for linear regression with *estar* embedding preference as the dependent variable and corpus as the predictor. As shown, all three corpora display a preference for *ser* embeddings with adjectives of evaluation (see Figure 4.3 for a visual representation of these results). Differently from adjectives of age and size, the Mexico corpus does not present *estar* extension with adjectives of evaluation. All three corpora, in fact, show the same patterns of copula use with this type of adjective (95% CIs [-0.02, 0.00] for both Arizona and Mexico, and [-0.03, -0.01] for Spain). The results for Mexico are not significantly different from zero, however ($p > .05$).

Table 4.6: Linear regression results with 95% confidence intervals for evaluation adjectives across Spanish baseline oral corpora.

corpus	Estimate	CIs	Pr(> t)
Arizona	-0.01	[-0.02, 0.00]	0.03 *
Mexico	-0.01	[-0.02, 0.00]	0.07
Spain	-0.02	[-0.03, -0.01]	0.00 ***

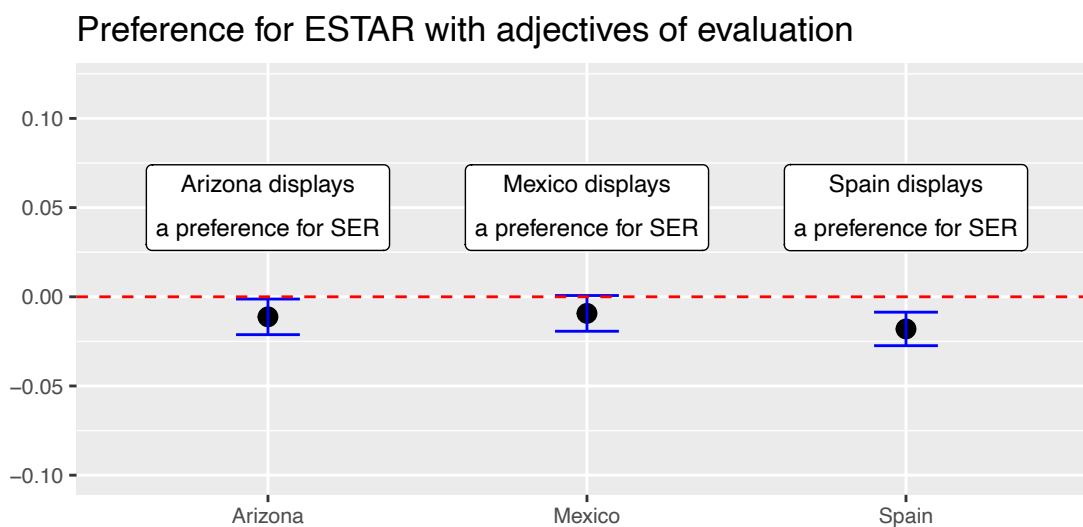


Figure 4.3: ESTAR Preference with adjectives of evaluation across Spanish baseline oral corpora. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

The utterances below are examples of copula use with different forms of the evaluation adjective *bueno/good* for the Mexico corpus (16-17), the Spain corpus (18-19), and the Arizona corpus (20-21). There are more instances of evaluation adjectives used with *estar* in the Mexico corpus (17).

(16) ya falleció pero **era buenísimo** (Mexico DF M23 024)

he already passed away but he was great

(17) lo que agarrábamos pues **estaba bueno** (Mexico Mexicali H31 027)

what we picked up was good

(18) aquello **era bueno** (Spain Madrid M33 054)

that was good

(19) no era, no **era bueno** (Spain Malaga H23 719)

it wasn't, it wasn't good

(20) eso **es buenísimo** (Arizona CESA 030)

this is great

(21) el trabajo no **es bueno** pero tengo trabajo (Arizona CESA 066)

the job is not good but at least I have a job

4.3 Conclusion

The goal of this first chapter was to answer the following research question:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?

Table 4.7 shows a summary of word embedding results for three Spanish corpora (i.e., Southern Arizona, Mexico, and Spain), displaying copula preference (i.e., *ser* or *estar*) by adjective type. Empty cells (-) mean results for that corpus and adjective type were not significant, i.e., that corpus and adjective type combination display no preference for either copula.

Table 4.7: Summary of word embedding results for the Spanish corpora.

Adjective Type	Arizona	Mexico	Spain
age	ser	estar	-
size	-	-	-
evaluation	ser	-	ser

As seen in the literature review, adjectives of age and size have been shown to favor *estar* in Spanish in Los Angeles (Cortés-Torres, 2004; Silva-Corvalán, 1986), New Mexico (Salazar, 2007), Sonora and Southern Arizona (Bessett, 2015). Results for *estar* embedding preference in this chapter show that while there is a significant difference across the three oral Spanish corpora in this dissertation, Arizona does not display a preference for *estar* embeddings with age, and size evaluation adjectives.

Evaluation adjectives have been shown to favor *estar* in Spanish in Los Angeles (Silva-Corvalán, 1986) and *ser* in Sonora (Bessett, 2015). However, in this dissertation, no significant differences were across the Spanish corpora, with all groups showing a preference for *ser* embeddings with adjectives of evaluation.

These results are likely to affect the analysis of the L3 Portuguese data, assuming learners' Spanish is most similar to the Arizona corpus here analyzed. Without this *estar* extension feature in the Arizona corpus, no contrasts are expected with Brazilian Portuguese. This is to be asserted in the L3 Portuguese results chapters. From now on, only the Southern Arizona corpus (i.e., CESA) is used as the baseline for Spanish as a potential source language for transfer.

CHAPTER 5

RESULTS 2 – EMBEDDINGS FOR BASELINE CORPORA

5.1 Introduction

This first results chapters addresses the following question:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?

I answer the question above in this chapter by discussing results of word embedding distances divided into two types. The first part of this chapter addresses *estar* embedding preference results across Arizonian Spanish and Brazilian Portuguese, a potential transfer source language and the target language respectively. The second part of this chapter discusses results of Portuguese *cognate* adjective embedding preference in American English.

As mentioned in the methods chapter, for word embeddings results, preference for *estar* across corpora is established based on the difference between word embeddings for *ser* and *estar* and the target lexical items in a predicate category (e.g., adjectives of size). Positive numbers indicate a preference for *estar*, while negative numbers, a preference for *ser*.

When discussing significant differences between corpora, I present statistical inference results for linear regression models for each type of predicate (e.g., size adjective, intensifiers), which were run with *estar* embedding preference as the dependent variable, corpus as independent variable, and word as a fixed effect (i.e., random intercept). The linear regression models are reported without contrasts (i.e., deviation from zero), with each level (i.e., corpus) presenting its own estimate mean (as opposed to having one factor level as the baseline, a.k.a. treatment contrasts) (Bolker, 2018; Fox & Monette, 2002). That is possible because all *estar* embedding preferences (i.e., the dependent variable) are centered on zero. The p values presented in

tables indicate whether that group mean is significantly different from zero (i.e., whether there is a preference for *estar* or *ser*). Difference between groups is established through confidence intervals. Results are divided by types of predicates.

5.2 Southern Arizonian Spanish vs. Brazilian Portuguese

In this section, preference for *estar* embeddings is compared between one of the potential transfer source languages (i.e., Arizonian Spanish) and the target language (i.e., Brazilian Portuguese). For inferential statistics, the same procedure ensues: linear regression models for each type of predicate (e.g., size adjective, intensifiers) were run with *estar* embedding preference as the dependent variable, corpus as independent variable, and word as a fixed effect. Estimates and confidence intervals are reported. Tables also present p values that indicate whether estimates are significantly different from zero. Results are divided by type of predicate. Positive numbers in charts and tables indicate a preference for *estar* embeddings, while negative numbers, a preference for *ser* embeddings.

5.2.1 ESTAR Preference with Adjectives

5.2.1.1 Age Adjectives

Table 5.1 shows linear regression results for *estar* embedding preference as the dependent variable and corpus as the independent variable (or predictor). As I have already discussed in the previous section, the Arizona corpus displays a preference for *ser* embeddings with adjectives of age (95% CIs [-0.06, -0.01], $p < .05$). The Brazil corpus, however, presents a wide range of variability (95% CIs [-0.10, 0.06], $p > .05$) Figure 5.4 displays a visual representation of the information found in Table 5.1 The reasons for the high variability could be a result of the smaller corpus size, but as demonstrated later in this chapter, the Brazil corpus shows significant preferences for one of the copula with adjectives of physical appearance and with intensifiers. Another possible explanation for the wide variability of *estar* embedding preference with adjectives of age is related to the composition of these

corpora. CESA comprises of sociolinguistic interviews that are much more structured than the private conversations and public speech in C-ORAL-Brasil, with this latter corpus covering a much broader range of topics and situations of use.

Table 5.1: Linear regression results with 95% confidence intervals for adjectives of age across Arizona and Brazil oral corpora.

corpus	Estimate	CI	Pr(> t)
Arizona	-0.03	[-0.06, -0.01]	0.02 *
Brazil	-0.02	[-0.10, 0.06]	0.68

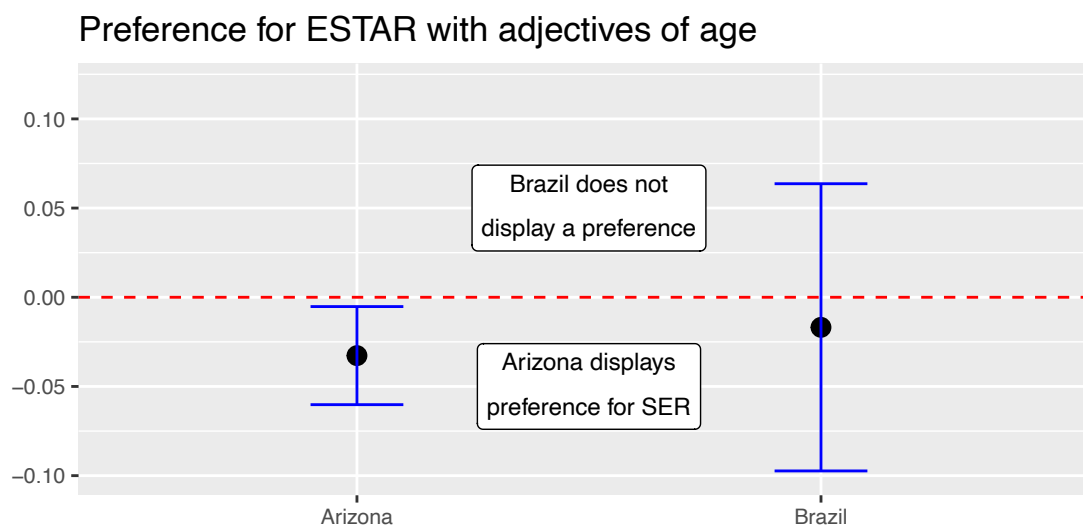


Figure 5.1: ESTAR embedding preference with adjectives of age across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

The examples below show the use of copula with age adjectives for the Arizona corpus (22-23) and the Brazil corpus (24-25). The Brazil corpus displays uses of both *ser* and *estar* with words like *velho/old*, while in the Arizona corpus adjectives of age like *joven/young* and *mayor/older* are used mainly with *ser*.

(22) cuando **era joven** mi hermana siempre salía (Arizona CESA 058)

when she was young my sister always went out

(23) aunque **era mayor** ella, ella siempre como lo protegía (Arizona CESA 020)

although she was older, she always protected him

(24) cigano é **velhaco** demais (Brazil C-ORAL bfamdl19)

the gypsy is too old

(25) cê **tá velha** (Brazil C-ORAL bfamcv15)

you are old

5.2.1.2 Size Adjectives

Table 5.2 presents linear regression results for *estar* embedding preference as the dependent variable and corpus as the independent variable (i.e., predictor). As already discussed, the Arizona corpus displays a preference for *ser* embeddings with adjectives of size, but these results do not differ significantly from zero (95% CIs [-0.03, 0.00], $p > .05$). The Brazil corpus again presents a wide range of variability and no preference for either copula embeddings with size adjectives (95% CIs [-0.07, 0.09], $p > .05$). For a visual representation of these results see Figure 5.2.

Table 5.2: Linear regression results with 95% confidence intervals for adjectives of size across Arizona and Brazil oral corpora.

corpus	Estimate	CI	Pr(> t)
Arizona	-0.02	[-0.03, 0.00]	0.11
Brazil	0.01	[-0.07, 0.09]	0.80

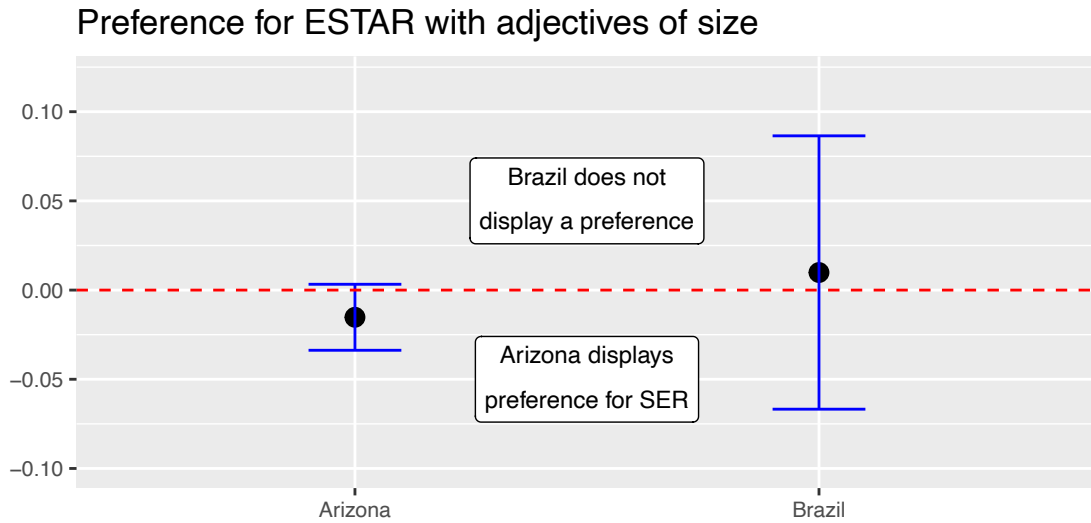


Figure 5.2: ESTAR embedding preference with adjectives of size across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

Both copula are used with adjectives of size in the Arizona corpus, mainly with the word *chico/little* (26-27). There is more variation in the Brazil corpus (28-29). While all forms of *pequeno/small* are used with *ser* in the Brazil corpus (28), *grande/big* is an adjective often used with both copula (29).

(26) pero me gusta porque es como, **es chico** (Arizona CESA 042)

but I like it because it is small

(27) cuando **estaba chica** sí, íbamos (Arizona CESA 013)

when I was little we went

(28) ele em si **é muito** pequenininho (Brazil C-ORAL bfamdl05)

he is very small

(29) **tá grande** demais (Brazil C-ORAL bfamcv11)

it is very big

5.2.1.3 Physical Appearance Adjectives

Table 5.3 presents linear regression results for *estar* embedding preference as the dependent variable and corpus as the independent variable (i.e., predictor). The Arizona corpus does not display a preference for either copula verb embeddings with adjectives of physical appearance (95% CIs [-0.02, 0.04] and $p > .05$). The Brazil corpus, however, displays a preference for *estar* embeddings with these adjectives (95% CIs [0.01, 0.11], $p < .05$). As seen in the literature review, the choice of *estar* with adjectives of physical appearance reflects speaker intention, and whether they are making a reference to a general vs. an individual norm (Falk, 1979; Woolsey, 2008). In other words, copula choice is dependent on whether the speaker means that the referent is beautiful or ugly in a comparison to a general norm of all people or to an individual norm. Figure 5.3 presents a visual representation of these results.

Table 5.3: Linear regression results with 95% confidence intervals for adjectives of physical appearance across Arizona and Brazil oral corpora.

corpus	Estimate	CI	Pr(> t)
Arizona	0.01	[-0.02, 0.04]	0.47
Brazil	0.06	[0.01, 0.11]	0.03 *

The examples below show both copula being used with adjectives of physical appearance in the Arizona corpus (30-32) and the Brazil corpus (33-34). In (32), for example, the speaker is talking about a young child that is crying, so the use of *estar* with *bonita/beautiful* is meant to say that although (or because) she was crying she was pretty (i.e., individual norm). In (33) the speaker is talking about the singer Madonna, so the use of *ser* means that they believe she is beautiful compared to other people (i.e. general norm). In (34) the speaker is talking about going to an event on the following day, so here he is referring to his individual norm.

(30) pero no está suficientemente grande para **estar muy feo** (Arizona CESA 021)

5.2.2 ESTAR Preference with Intensifiers (Adverb)

The presence of an intensifier (e.g., *muy*/very) modifying the head adjective in adjectival copula predicates has been shown to favor innovative *estar* use in some Spanish varieties (Salazar, 2007). Table 5.4 presents results for the linear regression with *estar* embedding preference as the dependent variable and corpus as the independent variable (i.e., predictor). As seen, the Arizona corpus does not display a preference for either copula embeddings with intensifiers (95% CIs [-0.02, 0.04], $p > .05$). The Brazilian Portuguese corpus, however, displays a preference for *ser* embeddings with intensifiers (95% CIs [-0.11, -0.01], $p < .05$). These results indicate that the Arizona corpus shows a slightly stronger tendency to prefer *estar* embeddings with intensifiers compared to the Brazil corpus (mean estimate for Arizona is at zero, and for Brazil it is at -0.06). For a visual representation of these results, see Figure 5.4.

Table 5.4: Linear regression results with 95% confidence intervals for intensifiers across Arizona and Brazil oral corpora.

corpus	Estimate	CIs	Pr(> t)
Arizona	0.00	[-0.01, 0.02]	0.46
Brazil	-0.06	[-0.11, -0.01]	0.03 *

The utterances below show the use of *estar* with intensifier in the Arizona corpus (35-36) and the Brazil corpus (37-38). Although both corpora seem to present both copula with intensifiers, uses of *ser* in the Brazil corpus as shown in (39) are overwhelmingly more frequent than *estar* uses (37-38).

(35) pero así **estaba muy divertido** (Arizona CESA 011)

but it was so fun

(36) al principio la llamada **estaba muy normal** (Arizona CESA 061)

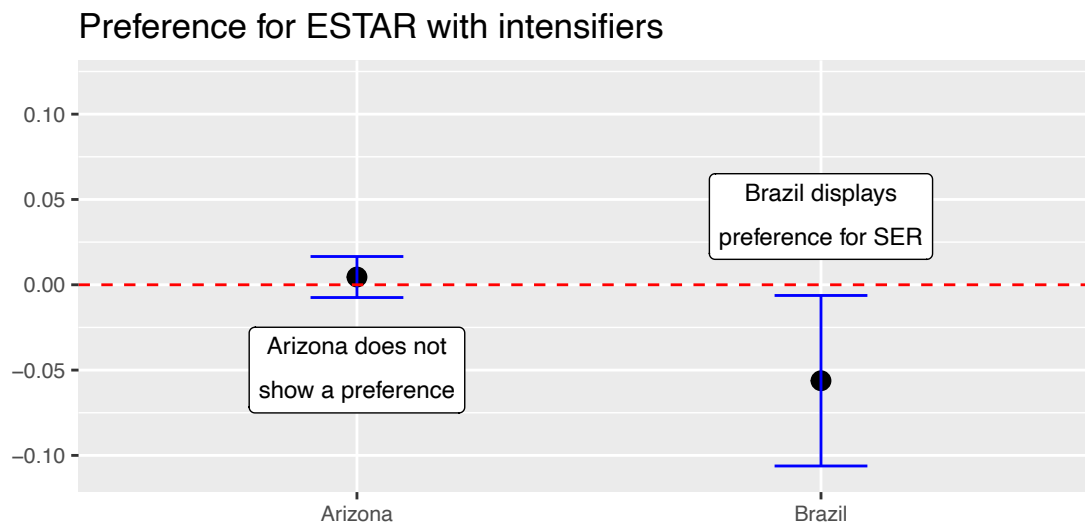


Figure 5.4: ESTAR embedding preference with EM prepositional predicates across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

at first the call was very normal

(37) ela **tá muito engraçada** (Brazil C-ORAL bfamdl26)

she was very funny

(38) e a Gabriela **tava muito teimosa** (Brazil C-ORAL bfamdl25)

and Gabriela was very stubborn

(39) a cidade **é muito boa** (Brazil C-ORAL bfammn24)

the city is really good

5.2.3 ESTAR Preference with EM/EN Prepositional Predicates

As described in the literature review, Portuguese presents more restrictions when it comes to using *estar* with *em/en* (in/at/on) prepositional predicates, with *ser* being used exclusively

in locatives with subjects that are immovable and *estar* with all other subjects. Table 5.5 presents linear regression for *em/en* prepositional predicates, with *estar* as the dependent variable and corpus as the independent variable. As shown, the Arizona corpus displays a clear preference for *estar* embeddings with *em/en* prepositions (95% CIs [0.23, 0.36], $p < .05$), matching the expectation that the Spanish corpus would prefer *estar* with *em/en* prepositional predicates (estimated mean of 0.29). The speakers in the Brazilian Portuguese corpus do not display a preference for either *ser* or *estar* copula embeddings, showing wide variability for *estar* embedding preference with *em/en* prepositional predicate (95% CIs [-0.35, 0.29], $p > .05$). That means the use of *estar* or *ser* with *em* prepositional predicates in the Brazilian Portuguese corpus is at chance (i.e., at 50% probability for the binary copula choice). Figure 5.5 presents a visual representation of these results, with the y axis showing a larger range than the other charts in this section due to the large variation present in the Brazil corpus for *estar* embedding preference linear results for *em/en* prepositional predicates.

Table 5.5: Linear regression results with 95% confidence intervals for EM prepositional predicates across Arizona and Brazil oral corpora.

corpus	Estimate	CIs	Pr(> t)
Arizona	0.29	[0.23, 0.36]	0.04 *
Brazil	-0.03	[-0.35, 0.29]	0.86

As mentioned, there are more restrictions to copula use with *em/en* prepositional predicates in Portuguese (43-44) than in Spanish (40-43). In Spanish, locatives are used with *estar* for referents that are inanimate and stationary like *escuela/school* (41-42). In Portuguese, only *ser* is used with inanimate and stationary references like *apartamento/apartment* (43). Referents that are animate or not stationary like *abuela/grandma* (40) and *carro/car* (44) are used with *estar* in *em/en* prepositional predicates in both languages.

(40) su abuela **estaba en** el México (Arizona CESA 008)

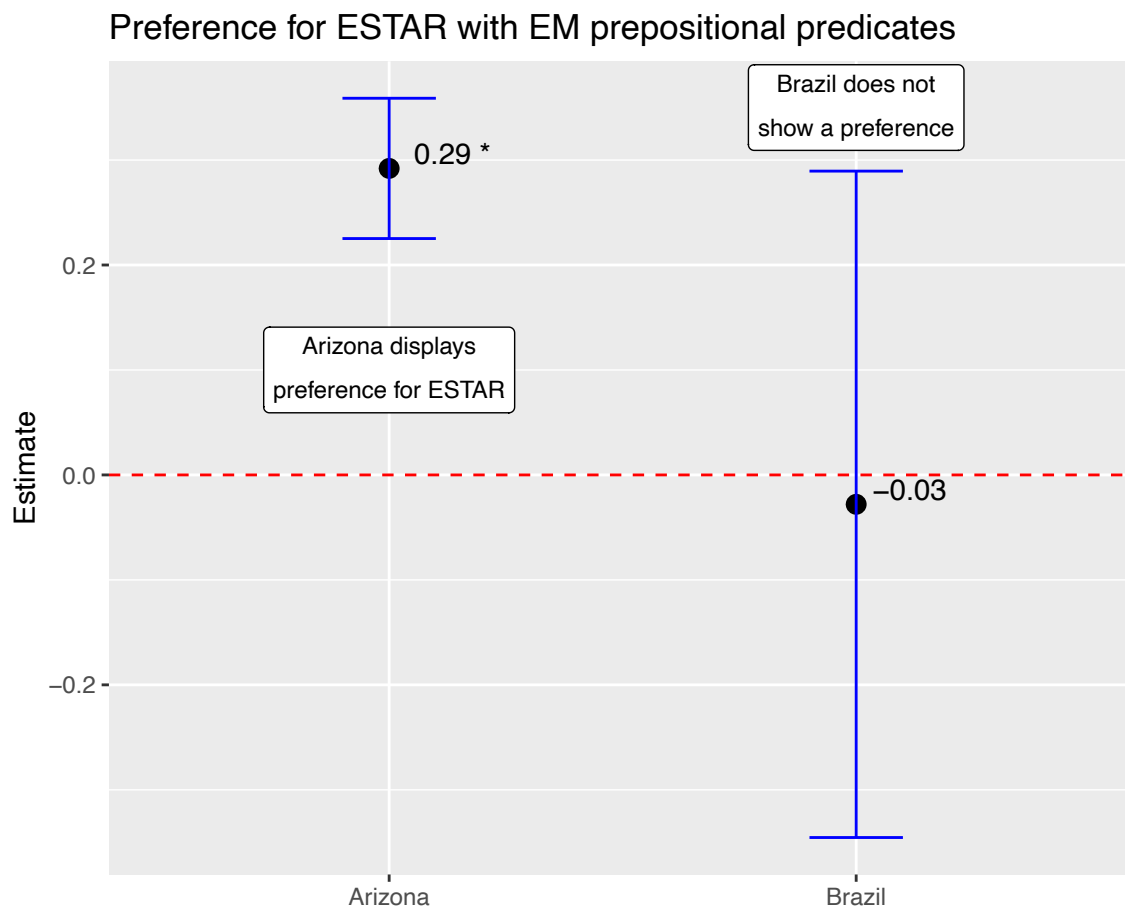


Figure 5.5: ESTAR embedding preference with EM prepositions across Southern Arizona and Brazilian Portuguese baseline oral corpora. Values above zero represent a preference for ESTAR embeddings, values below zero, a preference for SER embeddings. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

his grandma was in Mexico

(41) era XXX mi escuela, **estaba en** los Estados Unidos (Arizona CESA 047)

(it) was XXX my school, (it) was in the United States

(42) era una escuela muy iquita, **estaba en** las montañas (Arizona CESA 007)

(it) was a very small school, (it) was on the mountains

(43) o outro apartamento **era na** avenida praticamente (Brazil C-ORAL bfamcv29)

the other apartment was practically on the avenue

(44) o resto do carro já **estava no mato** (Brazil C-ORAL bfammn36)

the rest of the car was already in the woods

5.2.4 Summary of Differences: Arizona vs. Brazil

While adjectives of age and size favor *ser* in Spanish in Southern Arizona, the Brazilian Portuguese corpus displays a wide range of variation for these types of adjective. The Brazil corpus displays a wide range of variability for age and size adjectives, and also *em* prepositions, while the Arizona corpus shows a preference for *estar* embeddings. The Arizona corpus presents a clear preference for *estar* embeddings with *em* prepositions.

For adjectives of physical appearance, it is the Arizona corpus that does not display a preference for either copula embeddings, while the Brazil corpus shows a preference for *estar* embeddings. Regarding the presence of an intensifier in an adjectival predicate, the Arizona corpus does not seem to display a preference for either copula verb embeddings, while the Brazil corpus displays a clear preference for *ser* embeddings.

5.3 American English vs. Brazilian Portuguese

As discussed in the literature review, English has only one copula verb (i.e., *to be*) that maps over to *ser* and *estar* in Spanish and Portuguese. Since it is impossible to establish a preference for a certain copula *be* verb in English, the approach used to check whether English is a source of transfer into the L3 Portuguese production is through the use of cognates, and whether English cognate words are produced more often in copula predicate position in L3 Portuguese. The goal here is to establish whether there are any differences in adjective choice in copula predicate position for Portuguese cognate (which are usually latin-based words in English such as *special* and *intelligent*) versus non-cognate words across the different baseline English corpora.

Table 5.6 shows linear regression results for word embedding distance between *be* forms and adjectives as the dependent variable, cognate status as the independent variable, and word as a random effect (i.e., random intercept) for the CORE corpus. As expected, the distance between *be* forms and the various adjectives in the CORE corpus is non-zero ($p < .05$ for both cognate and non-cognate adjectives). While the difference in word embedding distances between cognate and non-cognate is not big (95% CIs [2.26, 2.38] and [2.31, 2.41] respectively), Portuguese cognate words seem to be slightly closer to *be* copula (estimate mean distance of 2.32) than non-cognate words (mean of 2.36). This slight preference for Portuguese cognate adjective embeddings might be due to mode of this corpus, which consists of written English texts found online.

Table 5.6: Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position of copula *be* in the CORE corpus.

cognate_status	Estimate	CI	Pr(> t)
cognate	2.32	[2.26, 2.38]	0 ***
non-cognate	2.36	[2.31, 2.41]	0 ***

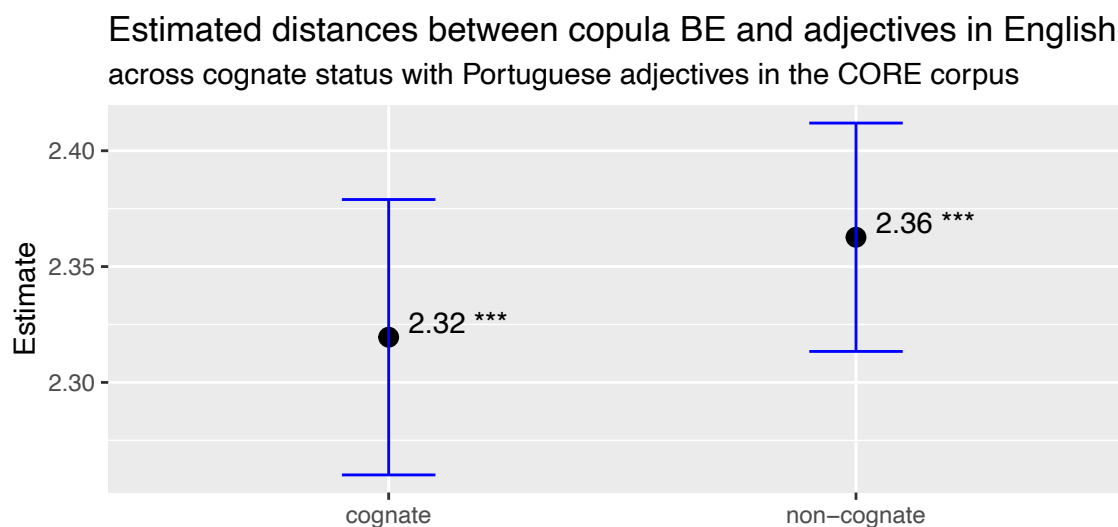


Figure 5.6: Estimated distance of adjectives in predicate position for English copula *BE* across cognate with Portuguese status for the CORE corpus. Blue bars represent 95% confidence intervals. Higher numbers represent bigger distances.

Table 5.7 shows linear regression results for word embedding distance between *be* forms and adjectives as the dependent variable, cognate status as the independent variable, and word as a random effect (i.e., random intercept) for the BangorTalk Miami corpus. Here again the distance between *be* forms and the various adjectives in the BangorTalk Miami corpus is non-zero ($p < .05$ for both cognate and non-cognate adjectives). Once again the difference in word embedding distances between cognate and non-cognate is not big (95% CIs [1.46, 1.70] and [1.42, 1.56] respectively), but this time Portuguese non-cognate words seem to be slightly closer to *be* copula (estimate mean distance of 1.49) than cognate words (mean of 1.58). The BangorTalk Miami corpus is a spoken corpus of English produced by L1 English speakers

who also speak Spanish as their first language. The mode of this corpus, i.e., speech, seems to affect the choice of adjective here, with non-cognate words being slightly preferred over Portuguese cognate words.

Table 5.7: Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position of copula *be* in the BangorTalk Miami corpus.

cognate_status	Estimate	CI	Pr(> t)
cognate	1.58	[1.46, 1.70]	0 ***
non-cognate	1.49	[1.42, 1.56]	0 ***

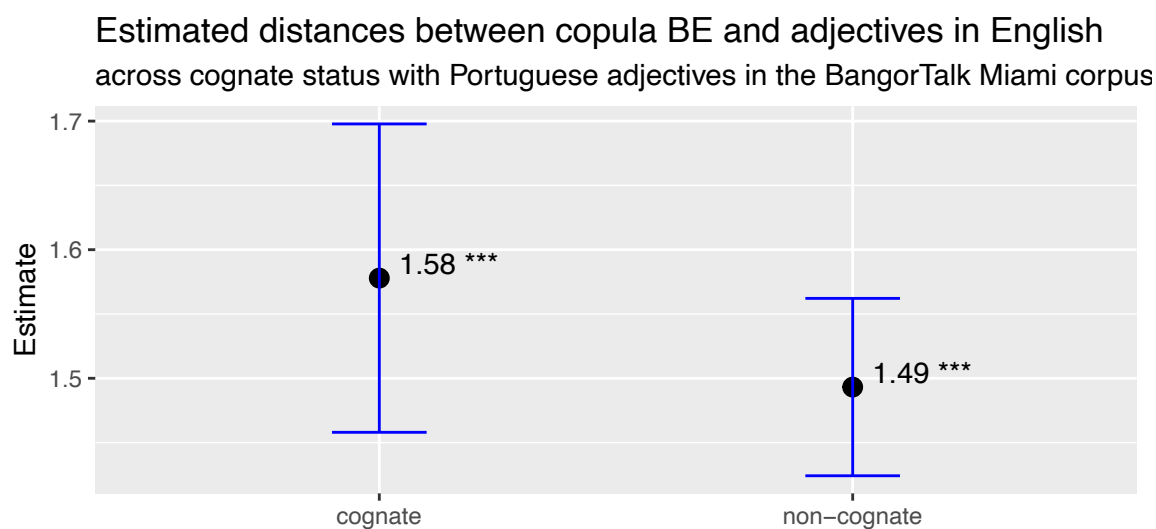


Figure 5.7: Estimated distance of adjectives in predicate position for English copula *BE* across cognate with Portuguese status for the BangorTalk Miami corpus. Blue bars represent 95% confidence intervals. Higher numbers represent bigger distances.

Table 5.8 shows linear regression results for word embedding distance between *be* forms and adjectives as the dependent variable, cognate status as the independent variable, and word as a random effect (i.e., random intercept) for the Cambridge corpus. As seen with the two previous English corpora, the distance between *be* forms and the various adjectives in the Cambridge corpus is non-zero ($p < .05$ for both cognate and non-cognate adjectives). The

95% confidence intervals for word embedding distances between cognate and non-cognate overlap completely (95% CIs [2.78, 3.00] and [2.84, 2.98] respectively). The Cambridge corpus is a very large spoken corpus of English produced by L1 English speakers across the United States. While the mode of this corpus is the same as the BangorTalk Miami corpus, i.e., speech, the Cambridge corpus might present more embedding variation due to its size.

Table 5.8: Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position of copula be in the Cambridge corpus.

cognate_status	Estimate	CI	Pr(> t)
cognate	2.89	[2.78, 3.00]	0 ***
non-cognate	2.91	[2.84, 2.98]	0 ***

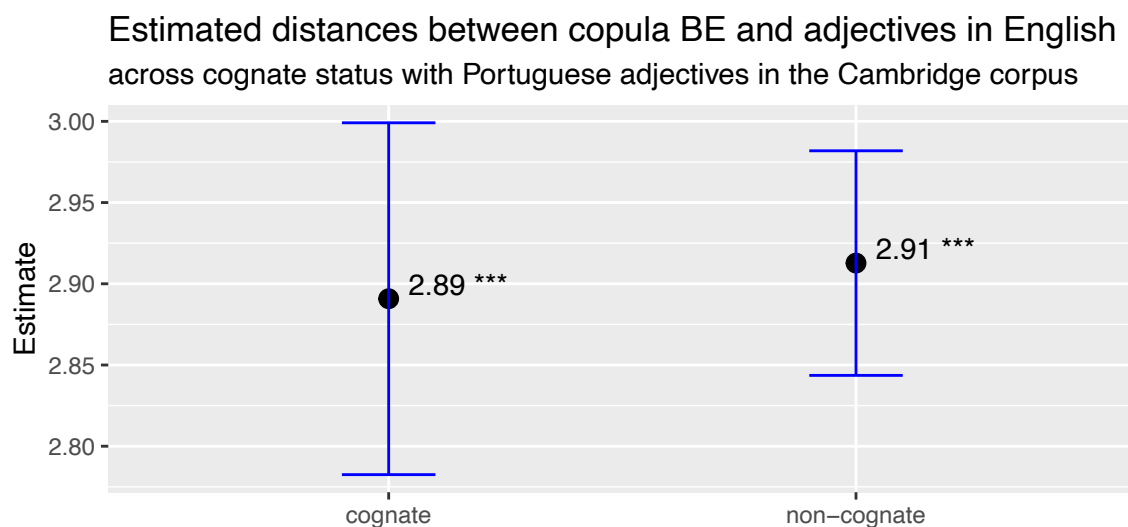


Figure 5.8: Estimated distance of adjectives in predicate position for English copula BE across cognate with Portuguese status for the Cambridge corpus. Blue bars represent 95% confidence intervals. Higher numbers represent bigger distances.

5.4 Conclusion

The goal of this first chapter was to answer the following research question:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?

Table 5.9 shows a summary of word embedding results for Southern Arizona Spanish and Brazilian Portuguese, displaying copula preference (i.e., *ser* or *estar*) by predicate type. Empty cells (-) mean results for that corpus and adjective type were not significant, i.e., that corpus and adjective type combination display no preference for either copula.

Table 5.9: Summary of word embedding results for base-line corpora.

Predicate Type	Arizona	Brazil
adjectival: age	ser	-
adjectival: size	-	-
adjectival: physical appearance	-	estar
intensifier presence	-	ser
prepositional: em	estar	-

Adjectives of age and size show no significant differences between the Southern Arizona and the Brazil corpora, due to high variability in the Brazil data. The Brazil corpus does show a preference for *estar* embeddings with adjectives of physical appearance, but here the Arizona corpus does not display a preference for either copula. In the chapter where I discuss results for the L3 Portuguese learner corpus at three different levels of proficiency, no differences are expected across proficiency levels for copula use with adjectives of age, size, and physical appearance.

Regarding the presence of an intensifier in an adjectival predicate, the Arizona corpus does not seem to display a preference for either copula verb, while the Brazil corpus displays a clear

preference for *ser* embeddings. For *em/en* prepositions, the Arizona corpus displays a strong preference for *estar* embeddings, while the Brazil corpus does not display any preferences. This last construction is of interest for the L3 Portuguese analysis because, if learners are transferring from Spanish to their Portuguese production, they will display a preference for *estar* with *em* prepositional predicates at lower levels of proficiency.

Regarding adjectives in predicate position in English copula (i.e., *to be*), it seems English adjectives that are Portuguese cognate are slightly preferred with copula *be* embeddings for the CORE corpus while non-cognates are preferred with copula *be* embeddings for the BangorTalk Miami corpus. The mode here (writing vs. speech) might explain these differences, with written production favoring Latin-based words (i.e. Portuguese cognate) such as *intelligent* over *smart*, with the opposite pattern produced in spoken production. This is also of interest for the L3 Portuguese analysis, which is based on a corpus of written texts. It is possible that if learners are transferring from English and relying on English cognates, there will be a stronger preference for English cognates in copula constructions at lower levels of proficiency. The mode (i.e., written) could affect these results. There is a lot of variation in the Cambridge corpus, and no differences are found across cognate versus non-cognate words. This might be a result of this corpus' size, since it is much larger than the other two (i.e., CORE and BangorTalk Miami). It has a greater number of speakers represented as well, which could be another factor in the larger variation in the embeddings for this corpus.

CHAPTER 6

RESULTS 3 – LOGISTIC REGRESSION FOR BASELINE CORPORA

6.1 Introduction

This results chapter addresses the following question:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?

To answer this question, each instance of copula use was extracted from the L1 Spanish and L1 Brazilian Portuguese corpora, and labeled as being an instance of *ser* or *estar* use for the Spanish and Portuguese corpora. Each copula construction was hand-checked for predicate type accuracy. Due to the need for this hand checking process, which is time consuming, and since the Spanish reference for this dissertation is Southern Arizona Spanish, only one Spanish corpus is included here as baseline (i.e., Arizona).

Logistic regression models were built with these binary variables as the dependent variable, corpus as the predictor (or independent variable), and participant was a random effect (i.e., random slope). Estimates from the logistic regression model, which are output as log odds, are then converted to probabilities.

The copula *estar* is used as the reference (i.e., coded as 1) for the L1 Spanish vs. L1 Portuguese results. Probabilities then reflect the probability of *estar* use for the copula constructions included in the regression model. Here no contrasts are used in the regression modeling, and significant p values (i.e., $p < .05$) mean that estimate probabilities are significantly different than pure chance (i.e., different from a 50% probability of *estar* use or *cognate* use).

Table 6.1 shows *estar* probability estimates for the two baseline corpora (i.e., L1 Brazilian Portuguese and L1 Arizonian Spanish) across the four types of copula predicates (i.e., adjective

phrase, adverbial phrase, noun phrase, and prepositional phrase). Significant differences between the two corpora are determined by 95% confidence intervals for the probability estimates (in square brackets in Table 6.1). As can be seen, there are no significant difference between L1 Portuguese and L1 Spanish when it comes to *estar* use with adjectival predicates (95% CIs [0.28, 0.35] and [0.30, 0.39] respectively), with both corpora selecting *estar* in these copula constructions at probabilities of .31 and .35 respectively (see Figure 6.1 for a visual representation of these results).

There does not seem to be significant difference between the two corpora when it comes to adverbial predicates either (95% CIs [0.50, 0.65] and [0.58, 0.73] respectively). However, *estar* selection with adverbial predicates in Spanish is at chance (i.e., not significantly different from .50, $p > .05$), while the results are significant for the Portuguese corpus (i.e., the Portuguese corpus shows a significant preference for *estar* with adverbial predicates). The results for noun phrase predicates show very low estimate probabilities for *estar* use with noun phrase predicates for both corpora (mean probability estimate of .07 for Spanish and .01 for Portuguese). As seen in the literature review, copula *ser* is expected to be used almost exclusively with this nominal predicates (Moura, 2016; Schmitt & Miller, 2007; Soschen, 2002) like in (1) and (2).

- (1) vocês **são o primeiro** (C-ORAL-BRASIL MIC)

you are the first

- (2) este **es la primera vez** que estoy hablando tanto español (CESA 001)

this is the first time that I am speaking so much Spanish

Exceptions for using *ser* with nominal copula predicates (i.e., use of *estar*) include colloquial expressions like (3) and (4) in Portuguese, and with words that mean child or kid like in (5) and (6) in Spanish:

- (3) ela **está gente boa** (C-ORAL-BRASIL DIN)

she is a good person

(4) **está uma porcaria** (C-ORAL-BRASIL TIT)

(it) is crap

(5) porque cuando **yo estaba niño** , ah , paseabamos en las bicicletas (CESA 006)

because when I was a kid, ah, we rode bikes

(6) cuando **estaba chamaca** que íbamos a la octrín y todo (CESA012)

when I was a child we would go to the Octrin and all that

With prepositional phrases, the L1 Portuguese corpus shows no preferences for either copula (mean = 0.50, 95% CIs [0.44, 0.57], $p > .05$). The L1 Spanish corpus differs significantly from the L1 Portuguese corpus here, showing a clear preference for *ser* with prepositional predicates (mean = 0.32, 95% CIs [0.28, 0.36], $p < .05$). These overall results illustrate how similar the two corpora are across the different types of copula predicate (see Figure 6.1 for this comparison), with most types of predicate following similar patterns for copula preference across the two corpora.

It is worthwhile to note that type of predicate (i.e., four levels: adjective phrase, adverbial phrase, noun phrase, and prepositional phrase) explains a lot of the variance in the data (R^2 fixed = 0.32). Individual differences, represented as participant as a random intercept in the model, do not explain that much more of the variance, for a total of 0.40 of variance explained for the entire model. When it comes to overall copula use, these participants behave as cohesive groups (i.e., small percentage of variance explained by individual differences).

Table 6.1: Logistic regression results with 95% confidence intervals across predicate types. First number under each L1 is the estimate mean ESTAR probability. Numbers inside the square brackets are 95% confidence intervals.

predicate_type	Spanish	Portuguese
AdjP	0.31, [0.28, 0.35], p = 0.00***	0.35, [0.30, 0.39], p = 0.00***
AdvP	0.57, [0.50, 0.65], p = 0.06	0.66, [0.58, 0.73], p = 0.00***
NP	0.07, [0.05, 0.08], p = 0.00***	0.01, [0.01, 0.02], p = 0.00***
PP	0.32, [0.28, 0.36], p = 0.00***	0.50, [0.44, 0.57], p = 0.93

6.2 Adjectival Predicates

The results in Table 6.1 above are omnibus results for each predicate type. In this section results are divided by semantic category. I focus my analysis of *estar* vs. *ser* use with adjectival predicates with descriptive, evaluation, and verbal adjectives. These are the types of adjectives that are also included in the L3 Portuguese analysis chapter, due to frequency of occurrence across all corpora analyzed in this dissertation.

6.2.1 Adjectives of Description

Description adjectives refer to nonmeasurable features of referents, and can be used with both *ser* and *estar* in Portuguese and Spanish. Table 6.2 presents probability estimates from logistic regression results for *estar* vs. *ser* use with these adjectives. As shown, both corpora present low probabilities for *estar* selection with description adjectives (mean estimate of 0.11 for Spanish and 0.25 for Portuguese). That means that constructions like (8) and (10) below are more common than (7) and (9). The L1 Arizonian Spanish corpus differs significantly from the L1 Brazilian Portuguese corpus (95% CIs [0.07, 0.16] and [0.17, 0.35], respectively),

ESTAR selection probability for different types of predicates in Spanish and Portuguese

R² fixed = 0.34 R² total = 0.4 | 9846 tokens and 237 participants

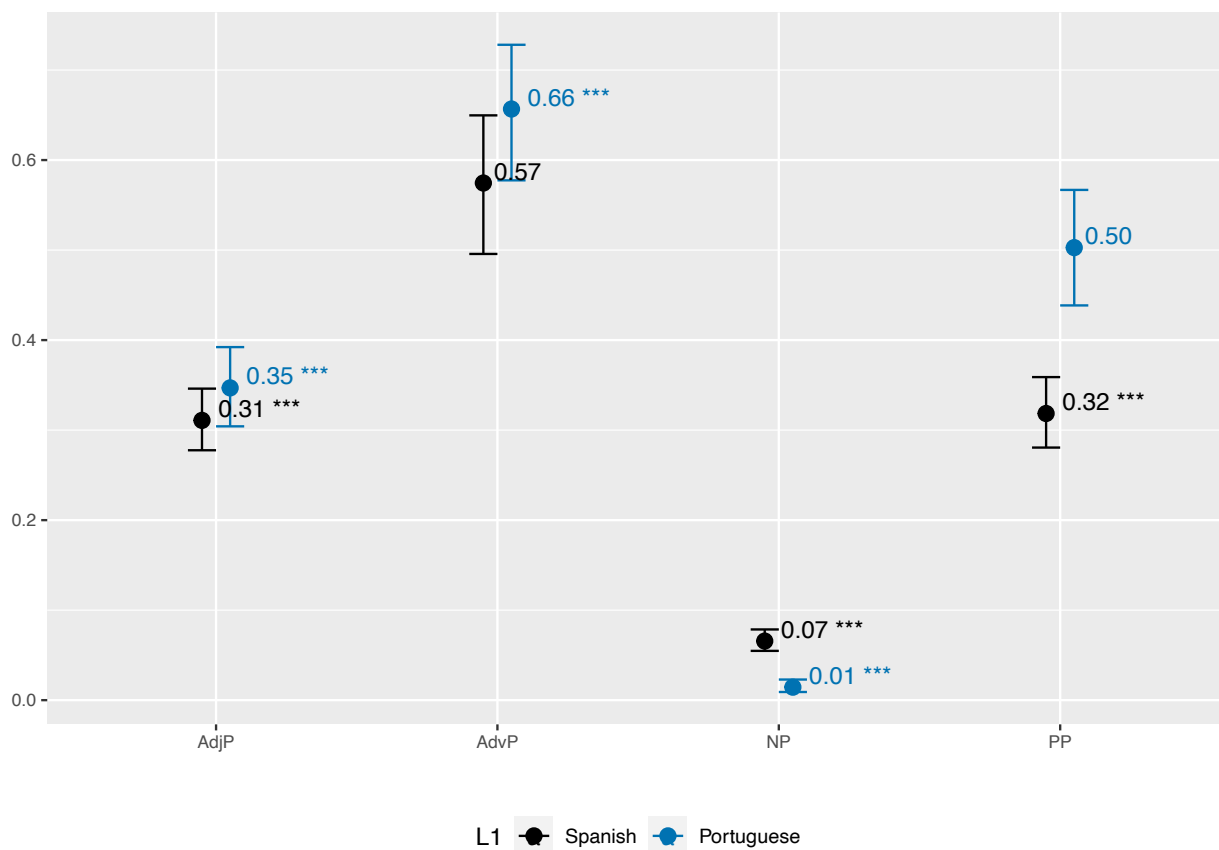


Figure 6.1: Logistic regression results (in probability estimates) for ESTAR vs. SER with 95% confidence intervals.

which is indicative of a stronger preference for *ser* with description adjectives in the Spanish baseline corpus. See Figure 6.2 for a visual representation of these results. The use of either *ser* or *estar* with description adjectives reflects speaker intent, and as such these results are affected by pragmatics that are not taken into account here. What is important to note is that L3 Portuguese learners are exposed to *estar* copula structures with adjectives of description in Portuguese, so it might be expected that they produce these constructions at non-zero probabilities.

Table 6.2: Logistic regression results for ESTAR vs. SER with description adjectives across L1s.

L1	Probability	CI	Pr(> z)
Portuguese	0.25	[0.17, 0.35]	0 ***
Spanish	0.11	[0.07, 0.16]	0 ***

(7) a casa **estava forte** (C-ORAL-BRASIL KAR)

the house is strong

(8) **é bem forte** (C-ORAL-BRASIL SAN)

(it) is really strong

(9) porque, cuando yo hablo en inglés, mi voz **está** muy **fuerte** (CESA007)

because, when I speak English, my voice is very strong

(10) la cultura latino aquí **es** más **fuerte** que en santa bárbara (CESA052)

the Latino culture here is much stronger than in Santa Bárbara

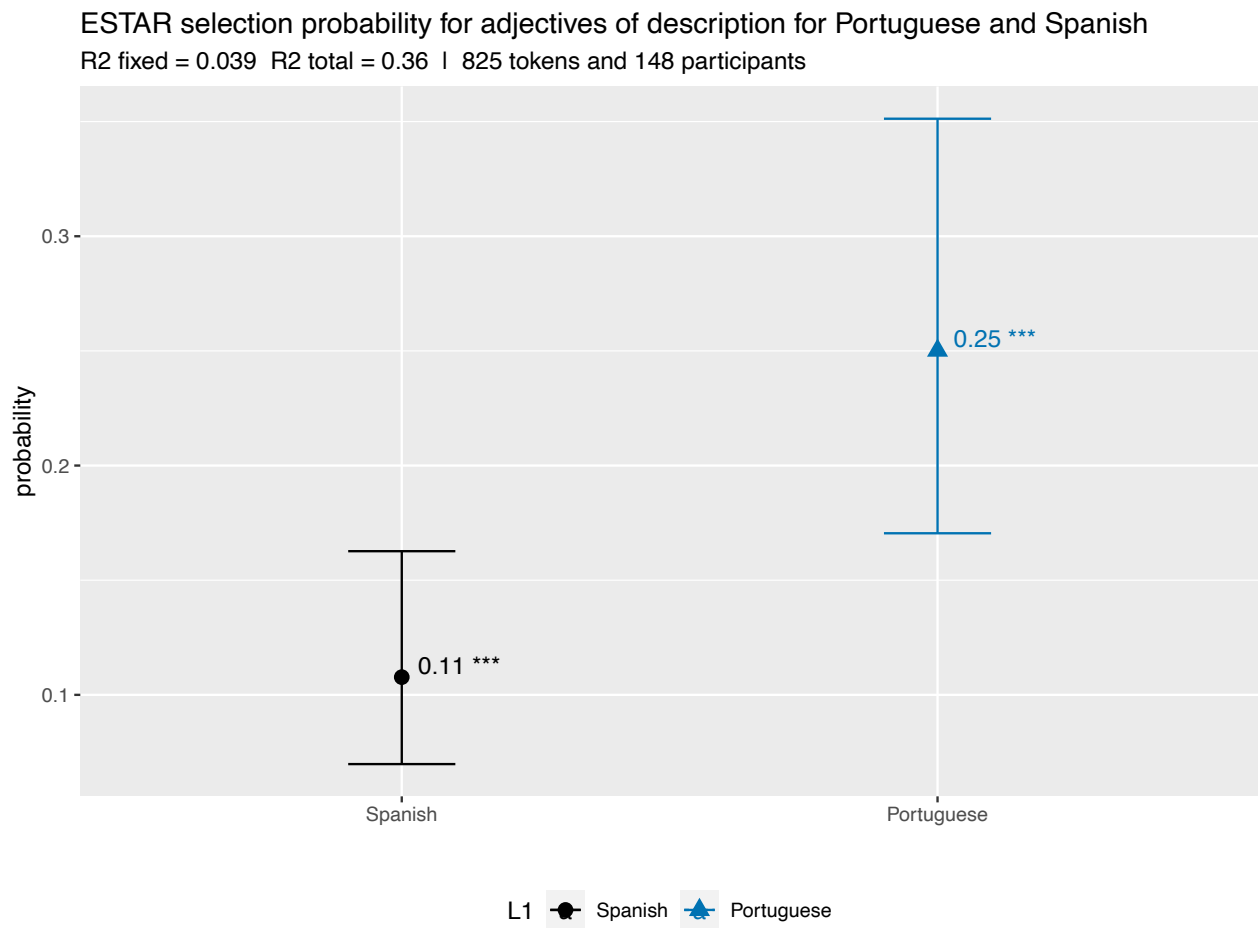


Figure 6.2: Logistic regression STAR selection probability estimates.

6.2.2 Adjectives of Evaluation

Adjectives of evaluation are also used with both *ser* and *estar* in Portuguese and Spanish. Table 6.3 displays estimate probabilities from logistic regression results for *estar* with evaluation adjectival predicates as the dependent variable and corpus as the predictor, with participant as a random intercept. Here again both corpora show a preference for *ser* with adjectives of evaluation (mean probability estimate of .05 for the Spanish corpus and .31 for the Portuguese corpus). That means that constructions like (12) and (14) below are more common than (11) and (13). The probability of *estar* selection with evaluation adjectives is much lower in the Spanish baseline corpus (95% CIs [0.03, 0.09]) compared to the Portuguese (95% CIs [0.21, 0.43]). The same issue of speaker intent is present here.

Table 6.3: Logistic regression results for ESTAR vs. SER with evaluation adjectives across L1s, with ESTAR as reference.

L1	Probability	CI	Pr(> z)
Portuguese	0.31	[0.21, 0.43]	0 **
Spanish	0.05	[0.03, 0.09]	0 ***

(11) outros lugares **estão bons** (C-ORAL-BRASIL EVN)

other places are good

(12) esse cara **é bom** também (C-ORAL-BRASIL LUC)

this guy is also good

(13) todos **están** muy **buenos** (CESA007)

everyone is very good

(14) pero, no **soy** muy **buena** (CESA047)

but, I'm not very good

ESTAR selection probability for adjectives of evaluation across corpora

R2 fixed = 0.15 R2 total = 0.48 | 779 tokens and 151 participants

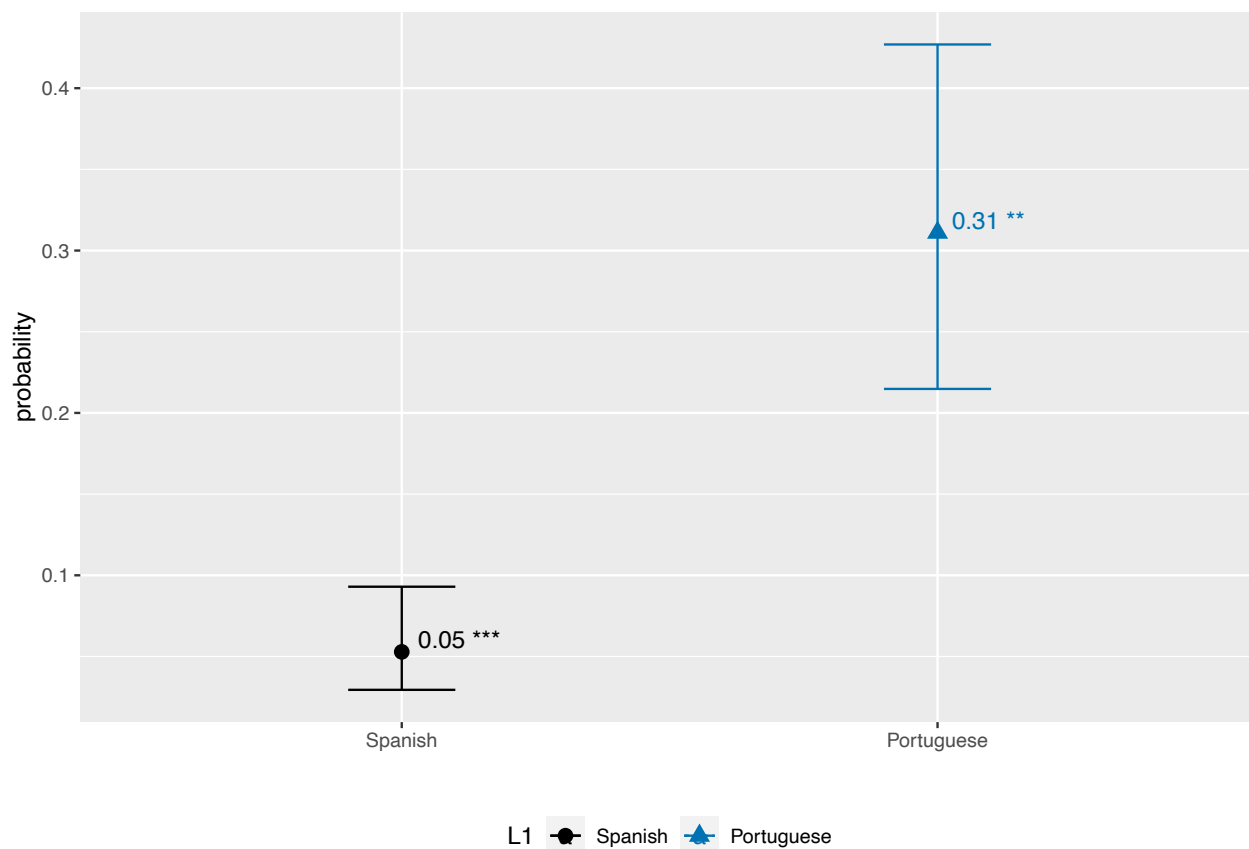


Figure 6.3: Logistic regression ESTAR selection probability estimates for evaluation adjectives.

6.2.3 Predicates with Verbal Adjectives

Verbal adjectives are mostly used with *estar* in both Portuguese and Spanish, since only the copula *estar* can be used with an adjective derived from an accomplishment verb (e.g., *abierto open*, *cansado tired*, and *confundido confused*) (Garavito & Valenzuela, 2006). Very few verbal adjectives can select *ser*, as in (16) and (18) below.

Table 6.4 shows the results for the logistic regression for *estar* vs. *ser* with corpus as a predictor and participants as a random intercept. As shown, the Portuguese and Spanish baseline corpora display the same probabilities for *estar* selection with verbal adjectives

(estimates = 0.85, 95% CIs [0.73, 0.92] and [0.77, 0.91] respectively). This is not surprising, since most verbal adjectives such as *arrepentido/regretful* and *cansado/tired* (11) are restricted to *estar* use in both languages. This is of interest, however, because there is a potential for transfer of these constructions from Spanish to L3 Portuguese.

Table 6.4: Logistic regression results for ESTAR vs. SER with verbal adjectives across L1s.

L1	Probability	CIs	Pr(> z)
Portuguese	0.85	[0.73, 0.92]	0 ***
Spanish	0.85	[0.77, 0.91]	0 ***

(15) você **está** **arrepentido** já (C-ORAL-BRASIL MAR)

you are sorry already

(16) isso daí **é** bem conhecido (C-ORAL-BRASIL ANT)

this is very well known

(17) **estoy** bien **cansada** todo el tiempo (CESA072)

I am very tired all the time

(18) pero **era** **avanzada** (CESA032)

but (it) was advanced

6.3 Prepositional EM Predicates

Portuguese has more restrictions when it comes to using *estar* with *em* (in/at/on) prepositional predicates, since *ser* is used with subjects that are immovable (20). The use *estar* with immovable subjects is grammatical in Spanish (21) but ungrammatical in Portuguese. With animate (e.g., a person) and inanimate subjects that can be moved (e.g., a small object like a

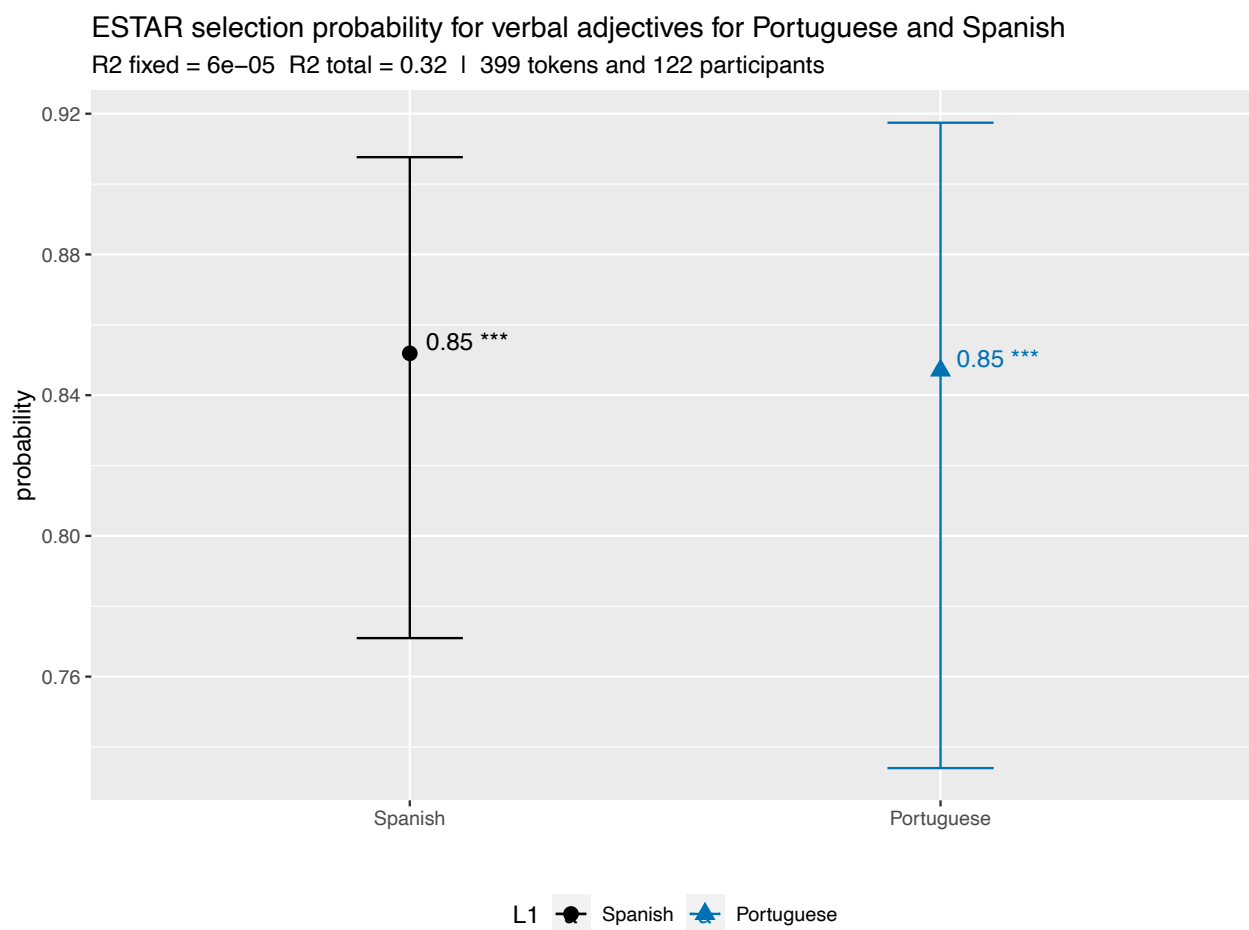


Figure 6.4: Logistic regression ESTAR selection probability estimates for verbal adjectives.

pen), *estar* is used in Portuguese (19). In addition to locatives, both Portuguese and Spanish use *ser* with other *em* prepositional predicates (22).

(19) **estava** lá **na** beira do rio (C-ORAL-BRASIL CAM)

(he) was there by the river (animate subject)

(20) Estúdio Bar **é** **no** centro (C-ORAL-BRASIL DAN)

Estúdio Bar is downtown (immovable subject)

(21) todas mis escuelas **estaban en** Arizona por eso aprendí inglés (CESA047)

all my schools were in Arizona that's why I learned English

(22) todo eso **eran en** inglés (CESA013)

all this was in English

Table 6.5 shows logistic regression results of corpus as a predictor of *estar* selection, and participants as a random intercept. As seen, the Spanish baseline corpus does indeed display a stronger probability of *estar* with *em* prepositional predicates (mean estimate of .89, 95% CIs [0.82, 0.93]) compared to the Portuguese baseline corpus (mean estimate of .64, 95% CIs [0.51, 0.76]).

Table 6.5: Logistic regression results for ESTAR vs. SER with EM prepositional predicates in Portuguese and Spanish, with ESTAR as reference.

L1	Probability	CIs	Pr(> z)
Portuguese	0.64	[0.51, 0.76]	0.03 *
Spanish	0.89	[0.82, 0.93]	0.00 ***

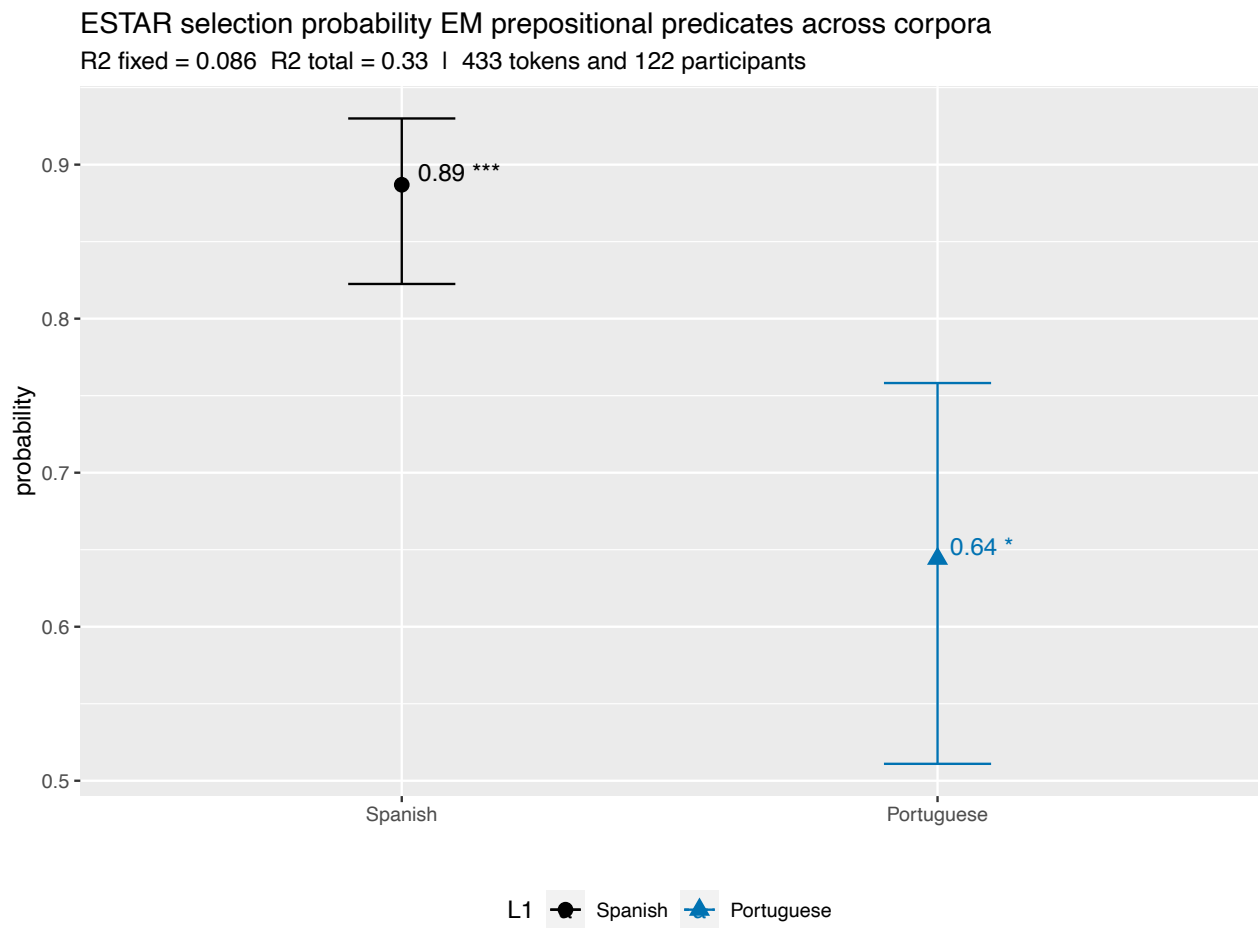


Figure 6.5: Logistic regression STAR selection probability estimates EM prepositional predicates.

6.4 Conclusion

Table 6.6 shows a summary of logistic regression results for Southern Arizona Spanish and Brazilian Portuguese, displaying copula preference (i.e., *ser* or *estar*) by predicate type. The estimated probability for *estar* selection is shown between parentheses.

Table 6.6: Summary of logistic regression results for base-line corpora. All results are significantly different from a chance probability.

Predicate Type	Arizona	Brazil
adjectival: description	<i>ser</i> (0.11)	<i>ser</i> (0.25)
adjectival: evaluation	<i>ser</i> (0.05)	<i>ser</i> (0.31)
adjectival: verbal	<i>estar</i> (0.85)	<i>estar</i> (0.85)
prepositional: em	<i>estar</i> (0.89)	<i>estar</i> (0.64)

This chapter first presented overall results of copula structures with all four types of predicate (i.e., adjective phrase, adverb phrase, noun phrase, and prepositional phrase) (Moro, 2000; Moura, 2016; Sibaldo, 2011). Overall, Arizonian Spanish (i.e., the Spanish baseline) and Brazilian Portuguese (i.e., the target language baseline), present similar patterns of copula preference across these predicate types. Both languages prefer *ser* with adjectival predicates. In addition, a much stronger preference is displayed for *ser* with noun phrase predicates for both languages, with a few exceptions mainly with slang expressions.

Overall results also show that while Brazilian Portuguese displays a preference for *estar* with adverbial predicates, Arizonian Spanish does not show a preference for either copula in these constructions. On the other hand, the Arizonian Spanish corpus shows a preference for *ser* with prepositional phrases overall, while Brazilian Portuguese does not display a preference for either copula.

Adjectival predicates split by type show that copula preference varies depending on the

specific type of adjective used. The specific adjective types presented in this chapter have been chosen due to their high frequency in the baseline corpora and the L3 Portuguese corpus (more details about these frequencies can be found in Chapter 7). For both description and evaluation adjectives, Brazilian Portuguese shows a stronger preference for *estar* than the Arizonian Spanish corpus. As mentioned, copula choice with these adjectives indicates speakers' intent and whether that quality is meant to be permanent or temporary (Woolsey, 2008). Nevertheless, the results presented in this chapter indicate that input in the target language contains non-zero levels of *estar* with these adjectives. As such, L3 Portuguese learners are expected to use *estar* with some of these adjectives at non-zero levels as well at some point in their L3 development, especially considering the broad range of assignment topics.

Results for copula choice with verbal adjectives are also included here due to the high occurrence of *estar* with this adjective type in both languages. This has great potential for the study of transfer from Spanish onto Portuguese production because *estar* is the copula that is marked for aspect and language learners usually default to the unmarked choice, which is *ser*, at lower levels of proficiency (Finnemann, 1990; Garavito & Valenzuela, 2006; Schmitt, 1992; Schmitt et al., 2004). If learners transfer this structure from Spanish, *estar* will be the preferred choice at low levels of proficiency. Otherwise, learners will default to the unmarked choice *ser* at first and start using *estar* with verbal adjectives at higher proficiency levels.

Results of copula use with *em* prepositional predicates for the two baseline corpora (i.e., L1 Arizonian Spanish and L1 Brazilian Portuguese) are also not surprising, with the Spanish corpus displaying a much higher probability for *estar* selection with this type of predicate compared to Brazilian Portuguese. Portuguese presents restrictions of copula use with *em* locatives that are not present in Spanish (Moura, 2016). If L3 Portuguese learners transfer these constructions from Spanish, they will produce non-target-like copula + *em* locatives structures, which will result in a higher probability of *estar* choice with *em* locatives in L3 Portuguese than those found in L1 Brazilian Portuguese.

In the next two chapters I present results that answer the questions on possible transfer of certain copula constructions from Spanish onto L3 Portuguese production.

CHAPTER 7

RESULTS 4 – EMBEDDINGS FOR L3 PORTUGUESE CORPUS

7.1 Introduction

This results chapter addresses the following question:

2. How do patterns of L3 Portuguese production by Spanish-English bilinguals compare to the source languages (i.e., Spanish and/or English) versus the target language (i.e., Portuguese) (at each of the three levels of L3 proficiency)?

This question is answered through linear regression of word embeddings results. This chapter focuses on *estar* embedding preference with copula predicates that can be used with both *ser* and *estar* in Spanish and Portuguese (i.e., structures that allow variation), including *estar* embedding preference with certain adjective classes and *em* preposition. Due to the size of the subcorpora split by level and L1 background (half a million words in total size divided across three subcorpora or three bilingual groups), results initially split by level or L1 background presented no differences across either of these factors. Variability present in the L3 data also might have been a factor here. Word embeddings were then calculated over the entire corpus for *estar* preference instead of individually split L1 background vs. proficiency level corpora. Only the results for the L3 data combined (as one corpus for the word embedding calculations) are presented for *estar* preference, since the split corpus results do not present any significant differences across groups.

For *cognate* preference, I target adjectival predicates. The dependent variable for the linear regression is the Euclidean distance between each copula instance to each adjective, with an adjective fitting into one of only two categories (i.e., English cognate and non-cognate), these results are divided across L1 background and L3 Portuguese level.

7.1.1 ESTAR Preference with Adjectives

This section includes three types of adjectives: age, size, and physical appearance, which have been investigated in studies on Spanish spoken in the United States (Bessett, 2015; Cortés-Torres, 2004; Salazar, 2007; Silva-Corvalán, 1986).

7.1.1.1 Age Adjectives

Table 7.1 shows linear regression results with 95% confidence intervals for word embeddings preference for *estar* (i.e., difference between *ser* and *estar* distances) with adjectives of age. Values above zero represent a preference for *estar*, values below zero, a preference for *ser*. According to the table, the Spanish corpus displays a preference for *ser* with age adjectives, while both Brazilian Portuguese and L3 Portuguese corpora do not display a preference for either copula. As can be seen, the mean estimate for the L3 Portuguese data is higher (i.e., a preference for *estar*) compared to both Arizona and Brazilian Portuguese estimates for *estar* preference with age adjectives. The variability in the L3 Portuguese data is, however, much greater (95% CIs [-0.45, 0.48] for L3 Portuguese and [-0.10, 0.06] for Brazilian Portuguese) than the other two corpora. That difference in variability can be more easily observed in Figure 7.1, which visually displays the same information as Table 7.1. The blue bars in the figure represent the 95% confidence intervals. The red dotted line is at zero, for reference.

Table 7.1: Linear regression results with 95% confidence intervals for adjectives of age across baseline oral corpora and L3 Portuguese.

group	Estimate	CI	Pr(> t)
L3 Portuguese	0.02	[-0.45, 0.48]	0.95
Portuguese	-0.02	[-0.10, 0.06]	0.68
Spanish	-0.03	[-0.06, -0.01]	0.02 *

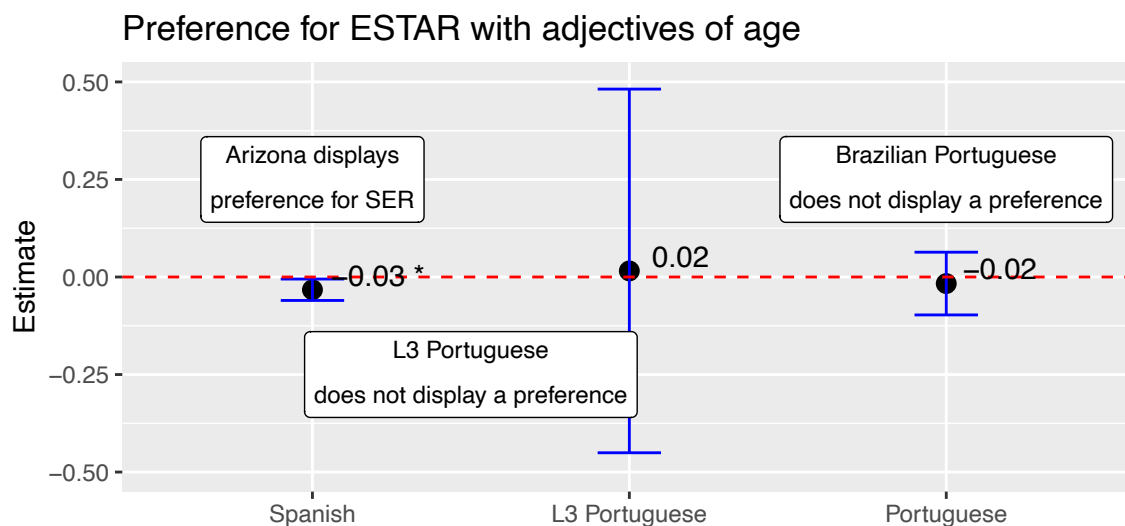


Figure 7.1: ESTAR Preference with age adjectives across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

7.1.1.2 Size Adjectives

Table 7.2 shows linear regression results with 95% confidence intervals for word embeddings preference for *estar* with adjectives of size. Values above zero represent a preference for *estar*, values below zero, a preference for *ser*. According to the table, the Spanish corpus displays a preference for *ser* with size adjectives, while both Brazilian Portuguese and L3 Portuguese corpora do not display a preference for either copula. As can be seen, the mean estimate for the L3 Portuguese is the same as the Brazilian Portuguese estimates for *estar* preference with age adjectives. The variability in the L3 Portuguese data is again greater (95% CIs [-0.35, 0.37] for L3 Portuguese and [-0.07, 0.09] for Brazilian Portuguese). For a visual comparison see Figure 7.2, which displays the same information as Table 7.2.

Table 7.2: Linear regression results with 95% confidence intervals for adjectives of size across baseline oral corpora and L3 Portuguese.

group	Estimate	CI	Pr(> t)
L3 Portuguese	0.01	[-0.35, 0.37]	0.95
Portuguese	0.01	[-0.07, 0.09]	0.82
Spanish	-0.02	[-0.04, 0.00]	0.04 *

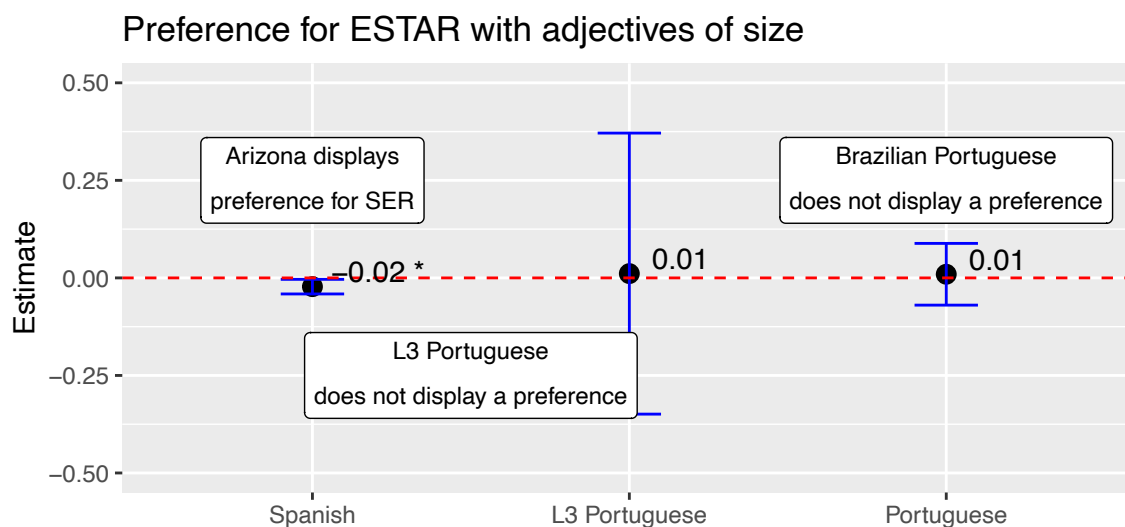


Figure 7.2: ESTAR Preference with size adjectives across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

7.1.1.3 Physical Appearance Adjectives

Table 7.3 shows linear regression results with 95% confidence intervals for word embeddings preference for *estar* with adjectives of physical appearance. Values above zero represent a preference for *estar*, values below zero, a preference for *ser*. According to the table, the Brazilian Portuguese corpus displays a preference for *estar* with physical appearance adjectives

(i.e., mean estimate and 95% confidence intervals are positive), while both Spanish and L3 Portuguese corpora do not display a preference for either copula. As can be seen, the mean estimate for the Spanish corpus shows a slight preference for *estar* with physical appearance adjectives, while the estimate for L3 Portuguese is right at zero. The variability in the L3 Portuguese data is here again the greatest of all three corpora (95% CIs [-0.36, 0.37] for L3 Portuguese, [0.01, 0.11] for Brazilian Portuguese, and [-0.02, 0.04] for Spanish). These comparisons are easier to observe in Figure 7.3, which displays the same information as Table 7.3.

Table 7.3: Linear regression results with 95% confidence intervals for adjectives of physical appearance across base-line oral corpora and L3 Portuguese.

group	Estimate	CIs	Pr(> t)
L3 Portuguese	0.00	[-0.36, 0.37]	0.98
Portuguese	0.06	[0.01, 0.11]	0.03 *
Spanish	0.01	[-0.02, 0.04]	0.47

7.1.1.4 English Cognate Adjectives

Table 7.4 displays linear regression results for *cognate* adjective preference represented by *copula + adjective* embedding distances as the dependent variable, level and bilingual group as the predictors, and word as random intercept. Negative values represent preference for *English non-cognate* adjectives, positive values a preference for *English cognate* adjectives. The L1 Spanish group is not included in these results due to their almost categorical use of *English non-cognate* adjectives. The word embedding calculation script only considers words that occur at least five times in the corpus, and for Levels 1 and 2 for the L1 Spanish group, there were no English-cognate adjectives that occur more than four times. As such, it was not possible to calculate the difference between English cognates and non-cognates (i.e., there were only non-cognates present in the word embeddings). Figure 7.4 shows the same

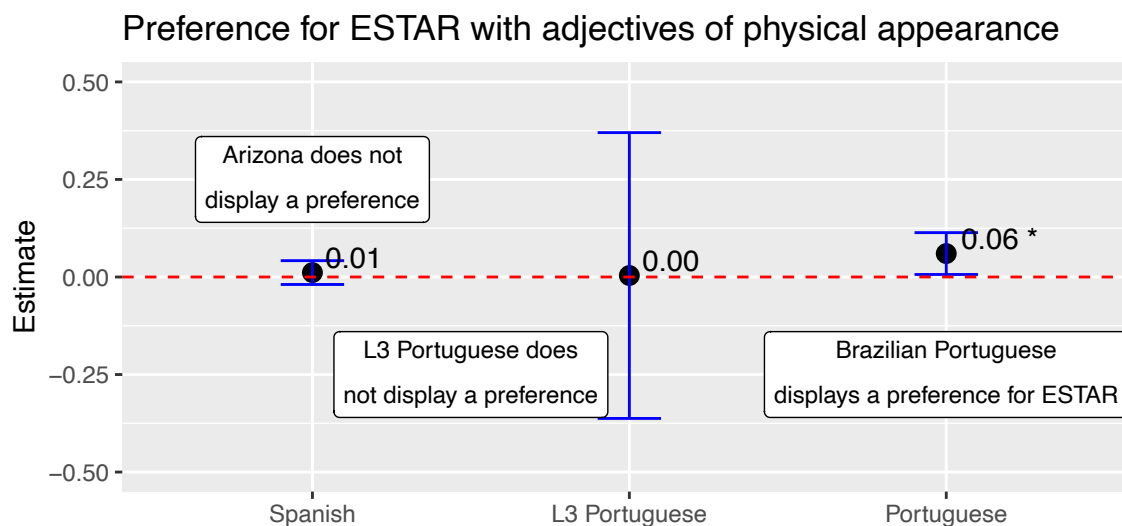


Figure 7.3: ESTAR Preference with physical appearance adjectives across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

results as Table 7.4, but in a visual manner. As can be seen, all groups have their estimate *cognate* preference under zero across all three L3 Portuguese levels (mean of -0.04 to -0.08 for the L1 English group, -0.12 to -0.23 for the Spanish Heritage group, and -0.08 for L1 Spanish). The L1 English group, however, displays a slightly stronger preference for *English cognate* adjectives than the Spanish Heritage group across all three L3 Portuguese levels. The preference for *English cognate* adjectives decreases from L3 Portuguese level 1 compared to level 3 for both L1 English and Spanish Heritage groups. No estimates for the first two levels are shown for L1 Spanish because these L3 Portuguese learners do have any cognate English adjective embeddings. That means the L1 Spanish group used Spanish cognate adjectives at a minimum frequency (word embeddings are calculated for tokens that show up at least 5 times in the data) for the first two L3 Portuguese levels.

Table 7.4: Linear regression results with 95% confidence intervals for cognate status for adjectives in predicate position.

L1	level	estimate	CI	
Spanish Heritage	level 1	-0.12	[0.07, -0.32]	***
Spanish Heritage	level 2	-0.29	[-0.07, -0.51]	***
Spanish Heritage	level 3	-0.23	[0.06, -0.52]	***
L1 English	level 1	-0.04	[0.15, -0.22]	*
L1 English	level 2	-0.07	[0.30, -0.43]	*
L1 English	level 3	-0.08	[0.17, -0.32]	***
L1 Spanish	level 3	-0.08	[0.19, -0.36]	**

Examples of most common English cognates include words like *natural* (1), *gay* (2), and *social* (3). These are also cognate with Spanish.

- (1) Não creio que deve ser controversa porque é natural (Level 1 - Heritage Spanish)

I don't believe that it should be controversial because it is natural

- (2) mas ninguém sabe que ele é gay (Level 2 - L1 English)

but nobody knows that he is gay

- (3) os meninos a vezes preferem de jogar na tecnologia em vez de ser social (Level 3 - L1 English)

the boys some times prefer to play with technology instead of being social

Examples of most common non-English cognates include words like *barato/cheap* (4), and *caro/expensive* (5). These are also cognate with Spanish.

- (4) A viagem na Havaí não é muito barato (Level 3 - L1 English)

The trip to Hawaii is not very cheap

(5) Os hotéis são muito caros (Level 3 - L1 Spanish)

The hotels are very expensive

English cognate preference for adjectives

R2 = 0.37

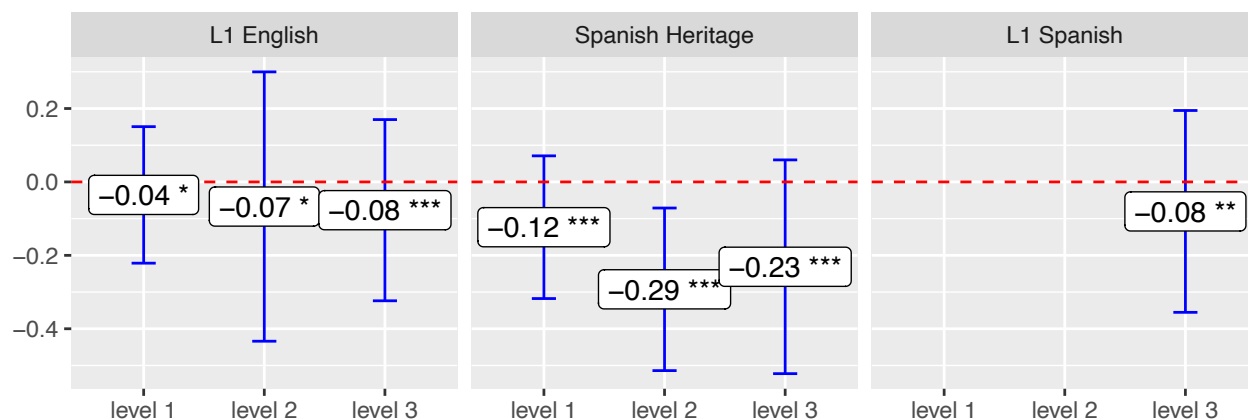


Figure 7.4: English COGNATE preference for adjectives across L1 English and Spanish Heritage groups. Values above zero represent a preference for English COGNATE, values below zero represent a preference for Spanish COGNATE. L1 Spanish group is not represented due to their exclusive preference for Spanish COGNATES.

7.1.2 ESTAR Preference with EM Prepositional Predicates

Portuguese presents restrictions when it comes to using *estar* with *em* (in/at/on) prepositional predicates, since *ser* is used with subjects that are immovable (7). L3 Portuguese learners often use *estar* instead (for non-target-like L3 uses see 8-9). With animate (e.g., a person) and inanimate subjects that can be moved (e.g., a small object like a pen), *estar* is used in Portuguese (6) which parallels the use in Spanish. In addition to locatives, both Portuguese and Spanish use *ser* with other *em* prepositional predicates (7-11).

(6) **estava** lá na beira do rio (C-ORAL-BRASIL CAM)

(he) was there by the river (animate subject)

- (7) Estúdio Bar **é** no centro (C-ORAL-BRASIL DAN)

Estúdio Bar is downtown (immovable subject)

- (8) *Natal **está** no Nordeste do Brasil (Level 1 - L1 English)

Natal is in Northeastern Brazil

- (9) *nosso hotel **estava** no centro da cidade (Level 3 - Heritage Spanish)

our hotel is in the center of the city

- (10) todas mis escuelas **estaban** en Arizona por eso aprendí inglés (CESA047)

all my schools were in Arizona that's why I learned English

- (11) todo eso **eran** en inglés (CESA013)

all this was in English

Table 7.5 shows linear regression results with 95% confidence intervals for word embeddings preference for *estar* with *em* prepositional predicates. Values above zero represent a preference for *estar*, values below zero, a preference for *ser*. According to the table, the Spanish corpus displays a preference for *estar* with *em* prepositional predicates (i.e., mean estimate and 95% confidence intervals are positive), while both Brazilian Portuguese and L3 Portuguese corpora do not display a preference for either copula. These results are not surprising because, as seen in the examples above, Spanish does use *estar* more broadly than Portuguese. As can be seen in Table 7.5, the mean estimate for the Brazilian Portuguese corpus shows a slight preference for *ser* with *em* prepositional predicates, while the estimate for L3 Portuguese is above zero, which indicates a preference for *estar* instead. These results might indicate transfer from Spanish, since L3 learners approximate Spanish patterns. However, the variability in the L3 Portuguese data is very large (95% CIs [-1.23, 1.47]) indicating greater instability of acquisition of this copula construction. These comparisons are easier to observe in Figure 7.5, which displays the same information as Table 7.5.

Table 7.5: Linear regression results with 95% confidence intervals for EM prepositional predicates across baseline oral corpora and L3 Portuguese.

group	Estimate	CI	Pr(> t)
L3 Portuguese	0.12	[-1.23, 1.47]	0.86
Portuguese	-0.03	[-0.35, 0.29]	0.86
Spanish	0.29	[0.23, 0.36]	0.04 *

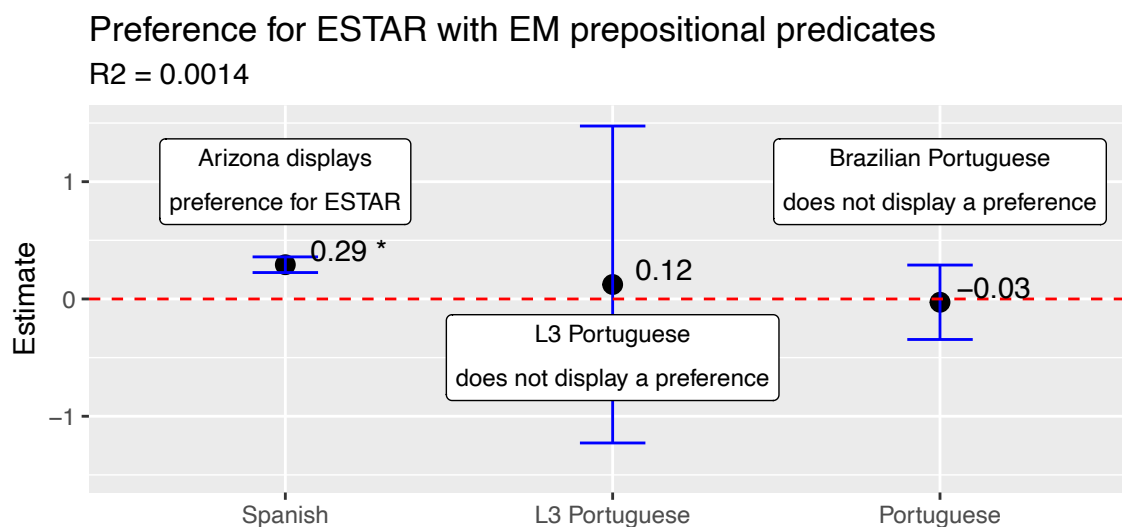


Figure 7.5: ESTAR Preference with EM prepositional predicates across baseline oral corpora and L3 Portuguese. Values above zero represent a preference for ESTAR, values below zero, a preference for SER. The red dotted line is at zero. Blue bars represent 95% confidence intervals.

7.2 Conclusion

Table 7.6 shows a summary of word embedding results for Southern Arizona Spanish, L3 Portuguese, and Brazilian Portuguese, displaying copula preference (i.e., *ser* or *estar*) by predicate type. Empty cells (-) mean results for that corpus and adjective type were not

significant, i.e., that corpus and predicate type combination display no preference for either copula.

Table 7.6: Summary of word embedding results for base-line and L3 Portuguese corpora.

Predicate Type	Arizona	L3_Portuguese	Brazil
adjectival: age	ser	-	-
adjectival: size	ser	-	-
adjectival: physical appearance	-	-	estar
prepositional: em	estar	-	-

It has been attested that word embeddings are more efficient with larger corpora (Goldberg & Levy, 2014; Mikolov et al., 2013; Zou et al., 2013), which is not the case when the L3 data in this dissertation is split into different subcorpora for L3 level or bilingual group. Splitting the data by L1 background and L3 level for *estar* embedding preference proved unproductive due to the low number of tokens per sub-corpus once the entire L3 corpus was divided and the linear regression modeling was run over a subset of word classes. As such, results for *estar* preference in this chapter only include embeddings there were calculated over the entire L3 corpus. Nevertheless, the results here presented suggest Spanish influence with *em* preposition: the L3 Portuguese data patterns more closely to Spanish, displaying more of a preference for *estar* with *em* preposition. Further analyses that split the learner data into the different L1 background groups and the three L3 proficiency levels are warranted. In the next chapter, results from logistic regression are presented, which allow this type of split for *estar* preference as a binary variable.

Regarding the preference of *English cognate* adjectives, the L1 Spanish group does not show up for the first two L3 Portuguese levels because of its categorical preference for *English non-cognate* words. In other words, the L1 Spanish group did not use enough English cognate tokens in the first two L3 Portuguese levels. For the two other groups, the cross-linguistic influence of English seems to be stronger for the L1 English group, which shows a slightly

stronger preference for *English cognate* words compared to the Spanish Heritage group. Across the three levels, both L1 English and Spanish Heritage groups show a decrease for *English cognate* words as level increases.

CHAPTER 8

RESULTS 5 – LOGISTIC REGRESSION FOR L3 PORTUGUESE CORPUS

8.1 Introduction

This results chapter addresses the following questions:

2. How do patterns of L3 Portuguese production by Spanish-English bilinguals compare to the source languages (i.e., Spanish and/or English) versus the target language (i.e., Portuguese) at each of the three levels of L3 proficiency?
 - a. Which language is the source of initial transfer for each Spanish-English bilingual group?
 - b. At what point (i.e., first, second or third semester) do L3 Portuguese patterns of copula use become most similar to L1 Portuguese (as opposed to most similar to L1 Spanish or L1 English)?
 - c. How similar are the L3 development paths across the three Spanish-English bilingual groups?

These questions are answered in this chapter through logistic regression results, where the binary variable is the selection of either *ser* or *estar* in copula structures. The models that included both bilingual group and L3 Portuguese level, with or without interaction, did not converge. Models that included linguistic factors such as tense-mood-aspect and referent type (e.g., pronoun, noun-phrase) also failed to converge. Regression models that fail to converge are unreliable and are thus not reported here. Therefore, only main effects and not interaction effects are reported. Results for two models are presented for most predicate types: a model with level as the predictor, and a model with L1 background as the predictor. Estimates for all models are output as log odds but are then converted and presented in probabilities, which are more human readable and indicate the probability of *estar* selection

for each group (either L3 Portuguese level or L1 background, and the baseline corpora). In addition, this chapter focuses on adjectival and prepositional predicates, which allow for both *estar* and *ser* use, as opposed to nominal predicates which are mainly used with *ser*.

Results for English cognates are also not included in this chapter due to the larger number of tokens that would be included in this type of analysis (8,652 total adjectives), and the need for hand-checking each individual copula construction. This will be done for future research.

8.2 Adjectival Predicate

I first focus my analysis of *estar* vs. *ser* use with adjectival predicates with adjectives that occur frequently in the corpora used in this dissertation.

Table 8.1: Token count across corpora and adjective type for adjectival predicates in copula structures.

Adjective Type	Total	Portuguese	L1 English	Heritage Spanish	L1 Spanish	Spanish
description	2423	218	615	716	267	607
evaluation	1810	235	398	472	161	544
miscellaneous	1339	69	439	474	161	196
verbal	1112	112	267	339	107	287
social class	472	17	50	45	55	305
size	376	42	44	55	26	209
emotional state	367	18	85	104	43	117
physical appearance	261	54	51	65	18	73
sensory	258	23	69	60	33	73
age	118	16	26	26	18	32
color	96	8	27	13	19	29
order	20	2	3	5	3	7
TOTAL	8652	814	2074	2374	911	2479

Table 8.1 shows total token counts of adjectival predicates by adjective type. I will present results for descriptive, evaluation, and verbal adjectives, whose totals are over one thousand tokens and individual corpus frequencies are over one hundred.

8.2.1 Adjectives of Description

As mentioned in the literature review and in previous results chapters, description adjectives (e.g., strong) can be used with both *ser* and *estar* in Portuguese and Spanish. The results for two logistic regression models are presented here, one with L3 level as the predictor of *estar* preference and another with L1 background as the L3 predictor instead. The unified model with both predictors did not converge, rendering the multivariate regression results unreliable.

Table 8.2: Variance explained by logistic regression model. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.

Regression Model	Fixed R2	Total R2
L3 level	0.085	0.32
L1 background	0.054	0.29

Table 8.2 displays the variance explained for each of these models. As seen, the logistic regression model with level as a predictor explains more of the variance in the data (R^2 fixed = .085) than the model with L1 background (R^2 fixed = 0.054), indicating that level is a slightly stronger predictor of *estar* selection with description adjectives. The difference between the fixed and the total variance explained in these models is large. The fixed variance explained expresses how much of the variance in the data is explained by the fixed effect (i.e., either L1 or level). The total variance explained includes both the fixed variance explained and the variance explained by the random effect (i.e., participant). The variance explained

rates in Table 8.2 indicate that individual variability is high and it explains a lot of the variance in the data. The estimate *estar* probabilities for each model are discussed in more detail below.

Table 8.3 displays estimated *estar* probabilities with 95% confidence intervals for logistic regression with L3 level and baseline corpora as the predictors and participant as a random effect (i.e., random intercept). As seen, Brazilian Portuguese displays a probability of 0.26 for *estar* selection with description adjectives, which is statistically different from Arizonian Spanish, which in turn displays a lower probability for *estar* selection (0.12). In addition, the results for L3 Portuguese across L3 level show that learners select *estar* with description adjectives at a much lower rate than the target language (i.e., Brazilian Portuguese) at the first two levels (mean *estar* probabilities of .05). At level 3, however, *estar* probability increases significantly (estimate = 0.15, 95% CIS [0.11, 0.20]), approximating the target language behavior (estimate = 0.26, 95% CIs [0.18, 0.35]). Figure 8.1 shows a visual representation of these results. The red line indicates the lower confidence interval for L1 Portuguese, which is only crossed by the L3 Portuguese data (in black) at the third L3 course level.

Table 8.3: Logistic regression results for ESTAR vs. SER with description adjectives across levels and baseline corpora.

Level	Probability	CIs	Pr(> z)
Portuguese	0.26	[0.18, 0.35]	0 ***
Level 1	0.05	[0.03, 0.07]	0 ***
Level 2	0.05	[0.03, 0.08]	0 ***
Level 3	0.15	[0.11, 0.20]	0 ***
Spanish	0.12	[0.08, 0.17]	0 ***

The L3 Portuguese examples below illustrate the preference for *ser* copula with description adjectives like in (1), which is stronger at Levels 1 and 2 (*estar* probability = .05). At Level 3, constructions with *estar* copula and description adjectives like (2) show an increase in

preference (*estar* probability = .15) compared to the two previous levels.

- (1) as cores *são* fortes e alegres (Level 2 - L1 Spanish)

the colors are strong and happy

- (2) muitas pessoas não realizam como o sol pode *estar* forte (Level 3 - Heritage Spanish)

many people don't realize how strong the sun can be

ESTAR selection probability for adjectives of description across levels

R2 fixed = 0.085 R2 total = 0.32 | 2423 tokens and 319 participants

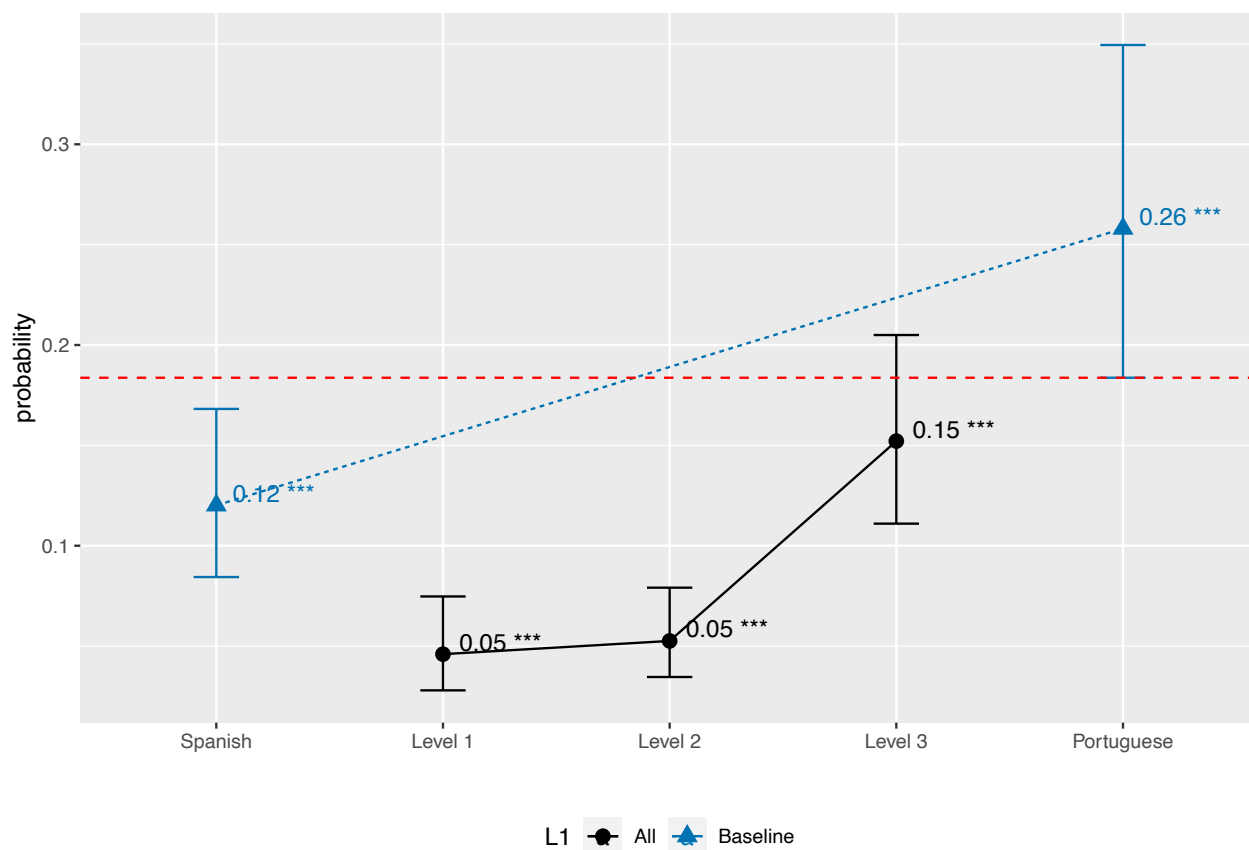


Figure 8.1: Logistic regression ESTAR selection probability estimates. The dotted red line marks lower confidence for Brazilian Portuguese, for reference.

Table 8.4 displays results for the model with L1 as a predictor. As seen, all learner groups display lower probabilities of *estar* selection with description adjectives, paralleling the Spanish baseline behavior. L1 English learners, however, show a stronger preference for the

unmarked copula (i.e., *ser*) (*estar* probability estimate = 0.06, 95% CIs [0.04, 0.09]). These probabilities are even lower than the Spanish baseline (estimate = 0.12, 95% CIs [0.09, 0.17]), indicating that L1 English learners are defaulting to *ser*. Figure 8.2 shows these results in a visual manner. The yellow dotted line marks the lower confidence interval for L1 Portuguese (i.e., the target language), while the black dotted line marks the lower confidence interval for L1 Spanish (i.e., transfer language). Note how the results for L1 English (in blue) are completely under the lower confidence interval for L1 Spanish, showing that L1 English chose *estar* with description adjective at lower probabilities than both baseline corpora (i.e., L1 Spanish and L1 Portuguese).

Table 8.4: Logistic regression results for ESTAR vs. SER with description adjectives across L1s and baseline corpora.

L1	Probability	CI	Pr(> z)
Portuguese	0.26	[0.18, 0.35]	0 ***
L1 English	0.06	[0.04, 0.09]	0 ***
Heritage Spanish	0.11	[0.08, 0.15]	0 ***
L1 Spanish	0.08	[0.04, 0.14]	0 ***
Spanish	0.12	[0.09, 0.17]	0 ***

8.2.2 Adjectives of Evaluation

Adjectives of evaluation (e.g., good) are also used with both *ser* and *estar* in Portuguese and Spanish. The results for one logistic regression model are presented for *estar* preference with evaluation adjectives, with L3 level as the predictor of *estar* preference. Both the model with L1 background as a predictor and the unified model with both L1 and level as predictors did not converge, rendering their results unreliable. Table 8.5 displays the variance explained for the model that converged, i.e., the logistic regression model with level as a predictor,

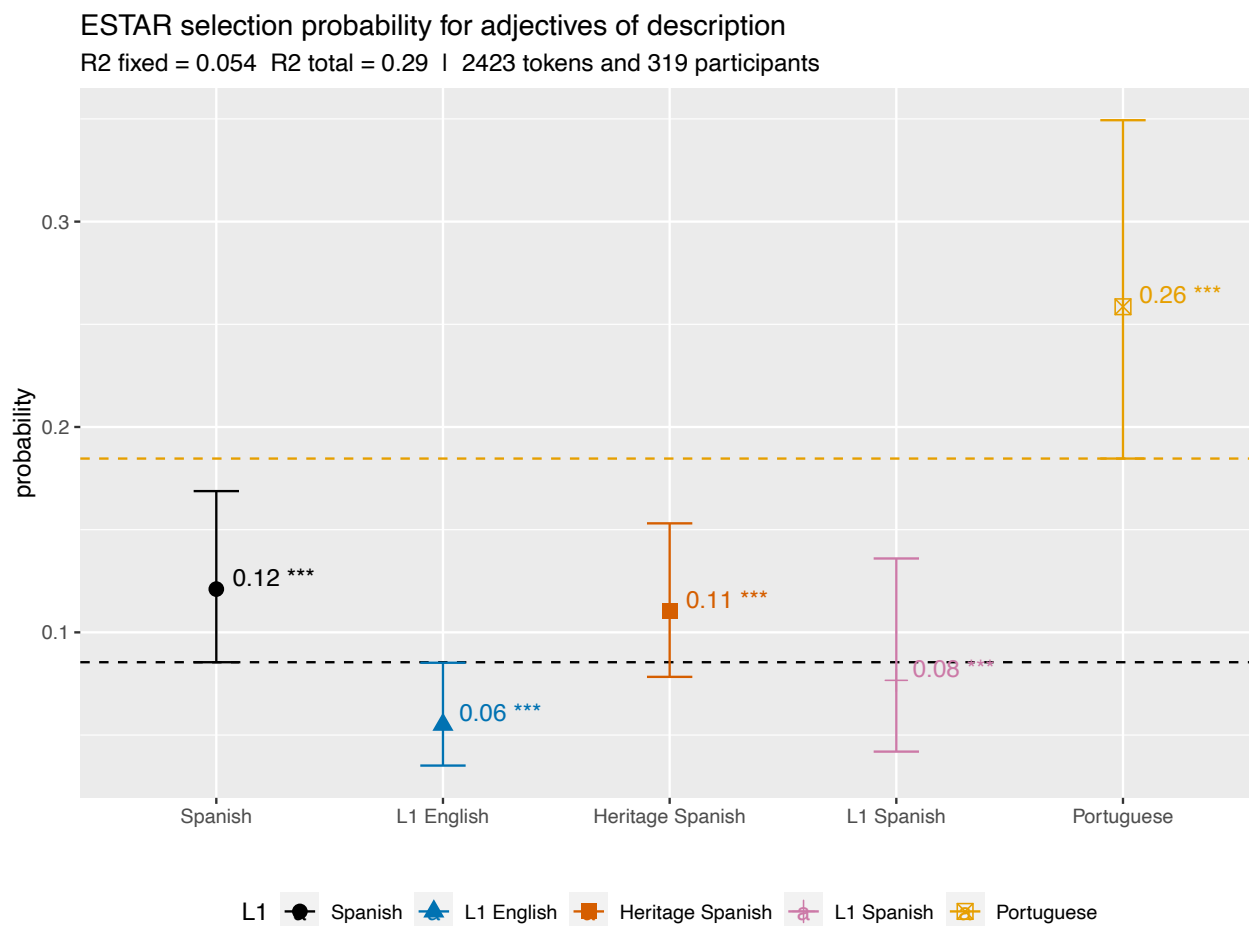


Figure 8.2: Logistic regression ESTAR selection probability estimates. The dotted yellow line marks lower confidence for Brazilian Portuguese and the black dotted line marks lower confidence for Spanish, for reference.

which explains a lot of the variance in the data (R^2 fixed = .29). This indicates that level is a strong predictor of *estar* selection with evaluation adjectives. The results for this model are discussed in more detail below.

Table 8.5: Variance explained by logistic regression model. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.

Regression Model	Fixed R2	Total R2
L3 level	0.29	0.58
L1 background	NA	NA

Table 8.6: Logistic regression results for ESTAR vs. SER with evaluation adjectives across levels and baseline corpora.

Level	Probability	CI	Pr(> z)
Portuguese	0.31	[0.21, 0.43]	0 **
Level 1	0.01	[0.00, 0.03]	0 ***
Level 2	0.01	[0.00, 0.03]	0 ***
Level 3	0.00	[0.00, 0.01]	0 ***
Spanish	0.05	[0.03, 0.09]	0 ***

Table 8.6 shows the regression model estimates for *estar* probabilities with 95% confidence intervals across baseline corpora and L3 level. As previously discussed, the probability of *estar* selection with evaluation adjectives is much lower in the Spanish baseline corpus (estimate = 0.05, 95% CIs [0.03, 0.09]) compared to the Portuguese baseline (estimate = 0.31, 95% CIs [0.21, 0.43]). As shown, the L3 learner corpus displays very low probabilities for *estar* selection across the three levels. Figure 8.3 show the baseline corpora in blue, with L1

Portuguese displaying higher *estar* probabilities than the Spanish baseline. The L3 data (in black) display very low *estar* probabilities, with confidence intervals all below the red dotted line, which marks the lower confidence interval for the L1 Spanish baseline corpus.

Copula constructions with *ser* and evaluation adjectives as in (3) are much more probable in the L3 data than copula *estar* with evaluation adjectives as in (4).

- (3) minha experiência na aula de português tem *sido* muito boa (Level 1 - L1 Spanish)

my experience in the Portuguese class has been very good

- (4) a música *está* muito boa (Level 2 - Heritage Spanish)

the music is very good

8.2.3 Verbal Adjectives

Verbal adjectives are mostly used with *estar* in both Portuguese and Spanish, since only the copula *estar* can be used with an adjective derived from an accomplishment verb (e.g., abierto *open*, cansado *tired*, and confundido *confused*). There are only a few verbal adjectives used with *ser*. The results for two logistic regression models are presented here, one with L3 level as the predictor of *estar* preference and another with L1 background as the L3 predictor instead. The unified model with both predictors did not converge, rendering the multivariate regression results unreliable. Table 8.7 displays the variance explained for each of these models. As seen, the logistic regression model with level as a predictor explains more of the variance in the data (R^2 fixed = .089) than the model with L1 background (R^2 fixed = 0.076), indicating that level is a stronger predictor of *estar* selection with verbal adjectives. The difference between the fixed and the total variance explained in these models is large, which is evidence that individual variability is high and it explains a lot of the variance in the data. The estimate *estar* probabilities for each model are discussed in more detail below.

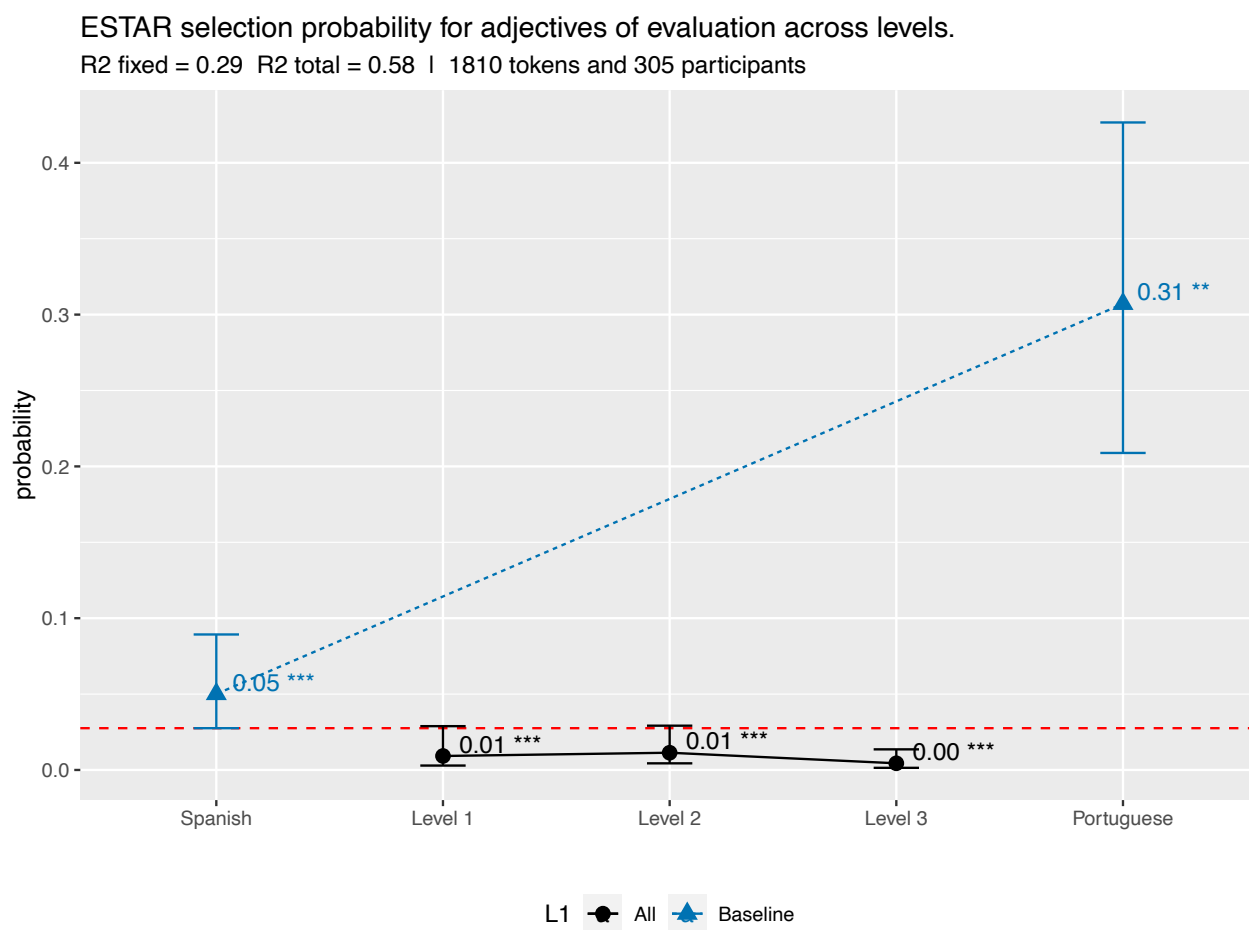


Figure 8.3: Logistic regression ESTAR selection probability estimates for evaluation adjectives. The dotted red line marks lower confidence for Spanish, for reference.

Table 8.7: Variance explained by logistic regression model for *estar* preference with verbal adjectives. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.

Regression Model	Fixed R2	Total R2
L3 level	0.089	0.39
L1 background	0.076	0.38

Table 8.8 displays the results for the logistic regression for *estar* preference with baseline corpora and L3 level as the independent variable, and individual as a random effect. As shown, the Portuguese and Spanish baseline corpora display the same probabilities for *estar* selection with verbal adjectives (estimates = 0.85, 95% CIs [0.74, 0.92] and [0.78, 0.91] respectively). Interestingly, L3 Portuguese learners at the first two levels show much lower probabilities for *estar* selection with verbal adjectives (estimates = 0.54 and 0.58, 95% CIs [0.43, 0.65] and [0.47, 0.68] respectively). In fact, at the first two L3 Portuguese levels, *estar* preference with verbal adjectives are at chance ($p > .05$). At level 3, however, the L3 Portuguese production is very similar to the target language (estimate = 0.83, 95% CIs [0.74, 0.89], $p < .05$).

Table 8.8: Logistic regression results for ESTAR vs. SER with verbal adjectives across levels and baseline corpora.

Level	Probability	CIs	Pr(> z)
Portuguese	0.85	[0.74, 0.92]	0.00 ***
Level 1	0.54	[0.43, 0.65]	0.45
Level 2	0.58	[0.47, 0.68]	0.13
Level 3	0.83	[0.74, 0.89]	0.00 ***
Spanish	0.85	[0.78, 0.91]	0.00 ***

L3 Portuguese copula constructions with verbal adjectives show a number of *estar* use as in (5) but also *ser* as in (6), which are both grammatical in Portuguese. However, (5) is more common in Brazilian Portuguese.

- (5) eu estava assustado na universidade (Level 3 - Spanish Heritage)

I was scared in the university

- (6) eu sou muito apaixonado pela arte (Level 2 - Spanish Heritage)

I am very passionate about art

Ungrammatical constructions with *ser* copula and verbal adjectives as in (7-9) occur across all three L3 Portuguese levels.

- (7) *Minha vida atualmente **é** muito ocupada (Level 1 - L1 English)

My life is currently very busy

- (8) *Alguns pessoas **são** preocupado que o espetáculo seria demais político (Level 2 - L1 English)

Some people are worried that the show would be too political

- (9) *Entendo que na sala de trauma **é** muito ocupado (Level 3 - Spanish Heritage)

I understand that in the trauma room it is very busy

Table 8.9 shows logistic regression results for *estar* preference with baseline corpora and L1 background as the independent variable, and individual as a random effect (i.e., random slope). As seen, *estar* probabilities across L1 background reveal that the L1 English speakers and the L1 Spanish speakers approximate the baseline corpora patterns (estimates = .72 and .80, 95% CIs [0.61, 0.81] and [0.64, 0.90] respectively). The Heritage Spanish speakers, however, do not display a clear copula preference with results for *estar* selection not being significant (estimate = 0.55, 95% CIs [0.45, 0.66], $p > .05$).

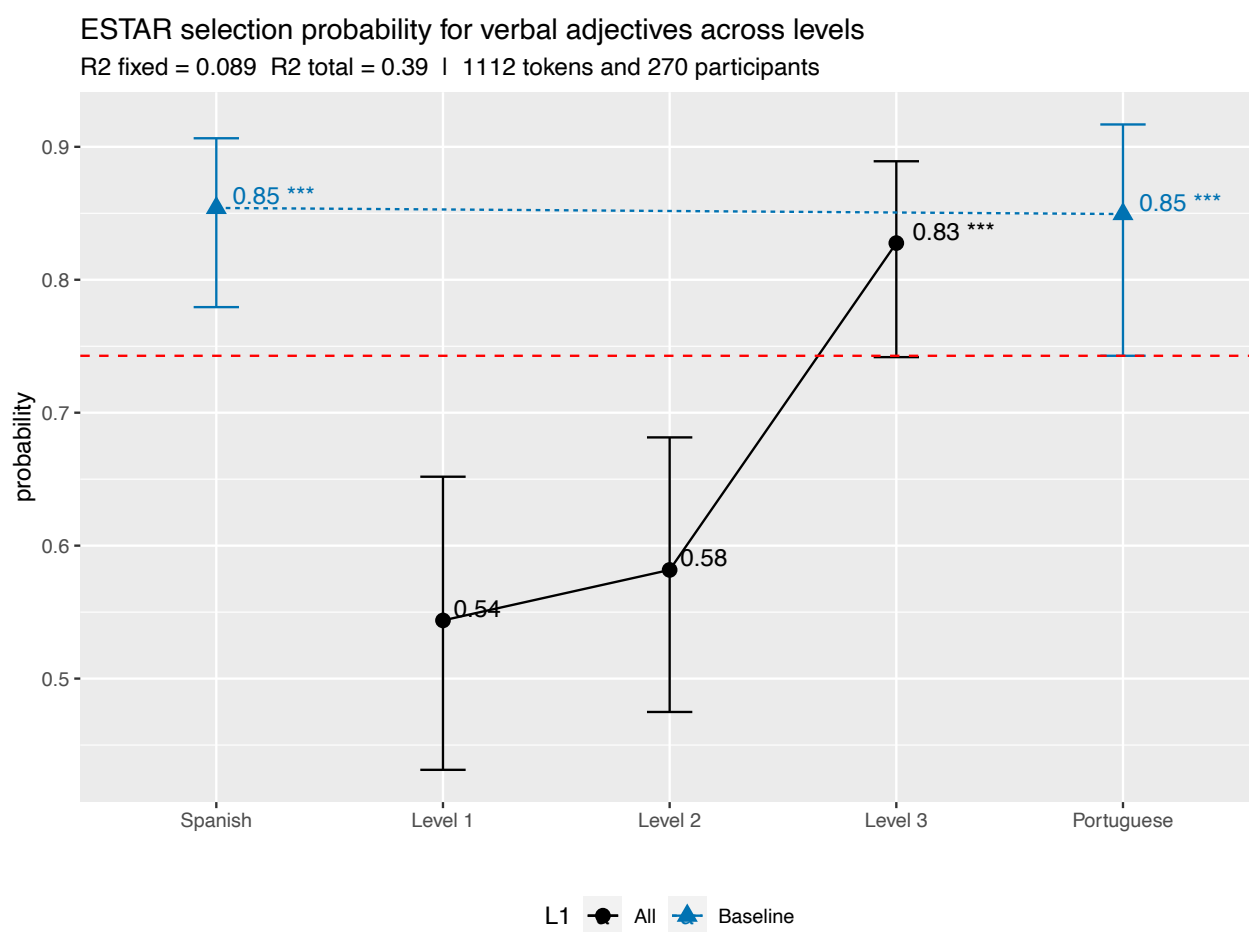


Figure 8.4: Logistic regression ESTAR selection probability estimates for verbal adjectives. The dotted red line marks lower confidence for Brazilian Portuguese, for reference.

Table 8.9: Logistic regression results for ESTAR vs. SER with verbal adjectives across L1 background and baseline corpora.

L1	Probability	CI	Pr(> z)	
Portuguese	0.85	[0.74, 0.92]	0.00	***
L1 English	0.72	[0.61, 0.81]	0.00	***
Heritage Spanish	0.55	[0.45, 0.66]	0.33	
L1 Spanish	0.80	[0.64, 0.90]	0.00	***
Spanish	0.85	[0.78, 0.91]	0.00	***

8.3 Prepositional EM Predicates

As previously discussed, Portuguese has more restrictions when it comes to using *estar* with *em* (in/at/on) prepositional predicates. In Portuguese, the *ser* copula is used in locative constructions with subjects that are immovable (e.g., a building), while *estar* is used in Spanish in these constructions.

Table 8.10: Total tokens and proportion of ESTAR tokens with EM preposition predicates across corpora.

L1	total	n	percent
Portuguese	114	71	0.62
L1 English	211	168	0.80
Heritage Spanish	223	160	0.72
L1 Spanish	93	56	0.60
Spanish	319	272	0.85
Total	960	727	0.76

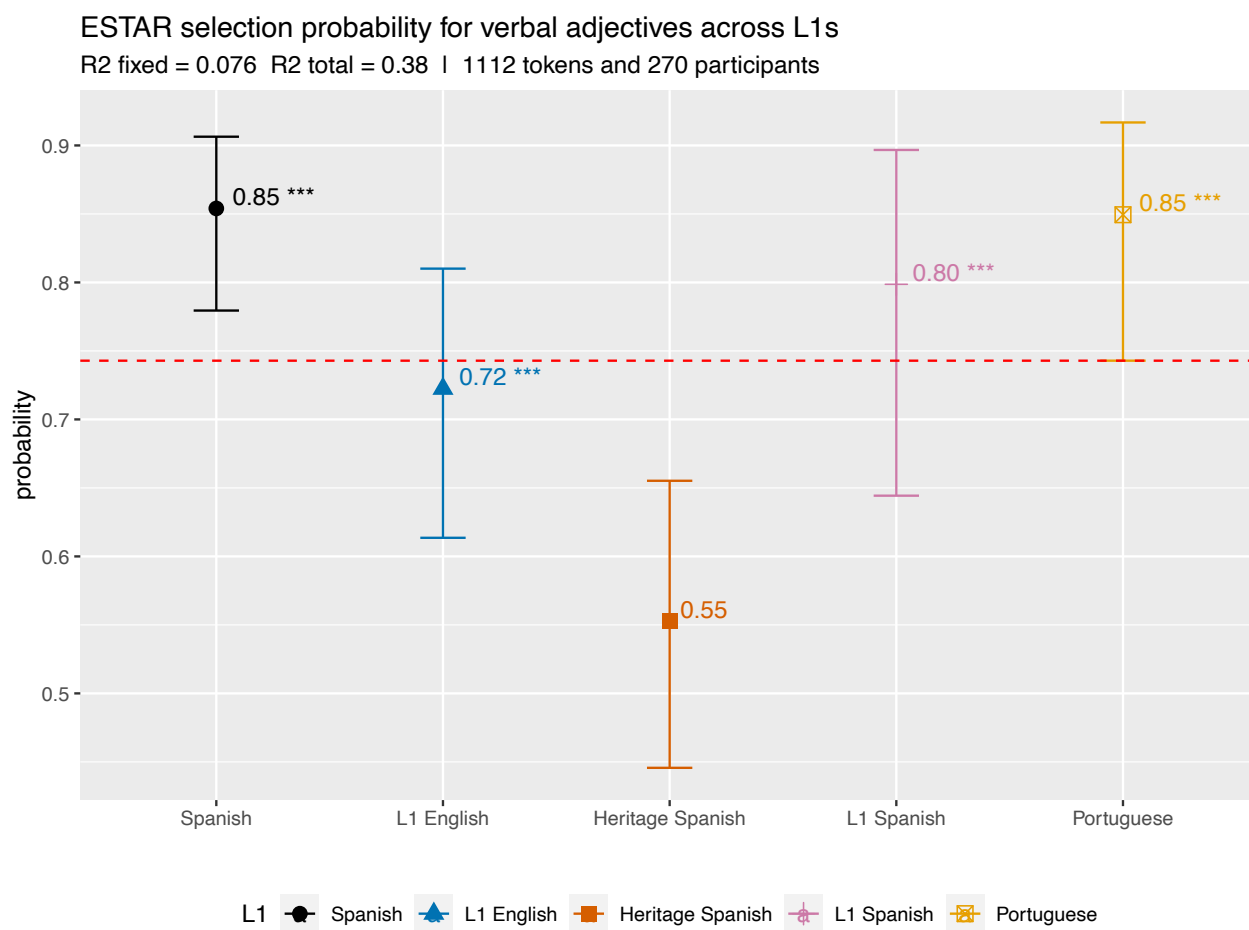


Figure 8.5: Logistic regression ESTAR selection probability estimates for verbal adjectives. The dotted red line marks lower confidence for Brazilian Portuguese, for reference.

Table 8.10 shows that *estar* frequency is lower for the L1 Portuguese corpus (.62) than for the L1 Spanish corpus (.85). The L3 Portuguese learners show a range of frequencies of *estar* copula with *em* prepositional (.60 to .85)

The results for two logistic regression models are presented for *estar* preference with *em* prepositional predicates: one with L3 level as the predictor of *estar* preference and another with L1 background as the L3 predictor. The unified model with both predictors did not converge, rendering the multivariate regression results unreliable. Table 8.11 displays the variance explained for each of these models. As shown, the logistic regression model with level as a predictor explains less of the variance in the data (R^2 fixed = .055) than the model with L1 background (R^2 fixed = 0.059), indicating that L1 background is a stronger predictor of *estar* selection with *em* prepositional predicate. Here again the difference between the fixed and the total variance explained in these models is large, which is evidence that individual variability is high and it explains a lot of the variance in the data.

Table 8.11: Variance explained by logistic regression model for *estar* preference with verbal adjectives. Fixed R^2 refers to the variance explained by fixed effects. Total R^2 refers to the variance explained for the whole model, including participant as a random effect.

Regression Model	Fixed R2	Total R2
L3 level	0.055	0.36
L1 background	0.059	0.36

Table 8.12 display the results for the logistic regression model with L3 level as a predictor of *estar* preference with *em* prepositional predicate. The Spanish baseline corpus displays a stronger probability of *estar* with *em* prepositional predicates (estimate = 0.90, 95% CIs [0.84, 0.94]) compared to the Portuguese baseline corpus (estimate = 0.90, 95% CIs [0.84, 0.94]). At levels 1 and 2, L3 Portuguese learners display *estar* selection probabilities that are in between the Spanish and Portuguese corpora (estimates = 0.81 and 0.84, 95% CIs

[0.70, 0.89] and [0.75, 0.90] respectively). At level 3 (estimate = 0.74, 95% CIs [0.62, 0.83]), however, L3 Portuguese learners approximate the target language behavior instead. Figure 8.6 shows these results with the baseline corpora in blue. The lower confidence interval for L1 Spanish and the upper lower confidence interval for L2 Portuguese are marked by the dotted red line. As seen, the L3 Portuguese data (in black) is in between the two baseline corpora in terms of *estar* probabilities with *em* prepositional predicates.

Table 8.12: Logistic regression results for ESTAR vs. SER with EM prepositional predicates across L3 Portuguese levels and baseline corpora.

Level	Probability	CIs	Pr(> z)	
Portuguese	0.65	[0.51, 0.77]	0.04	*
Level 1	0.81	[0.70, 0.89]	0.00	***
Level 2	0.84	[0.75, 0.90]	0.00	***
Level 3	0.74	[0.62, 0.83]	0.00	***
Spanish	0.90	[0.84, 0.94]	0.00	***

L3 Portuguese learners often use *estar* in *em* locative constructions as they would in Spanish, as in (10) and (11), which are ungrammatical in Portuguese.

(10) *Natal está no Nordeste do Brasil (Level 1 - L1 English)

Natal is in Northeastern Brazil

(11) *nosso hotel estava no centro da cidade (Level 3 - Heritage Spanish)

our hotel is in the center of the city

Table 8.13 displays logistic regression results for *estar* preference with *em* prepositional predicates with baseline corpora and L1 background for the L3 data as the predictor. As seen, the L1 English speakers display *estar* preference with *em* prepositional predicates that are similar to the Spanish baseline corpus (mean estimates of .86 and .89, 95% CIs [0.77, 0.91] and

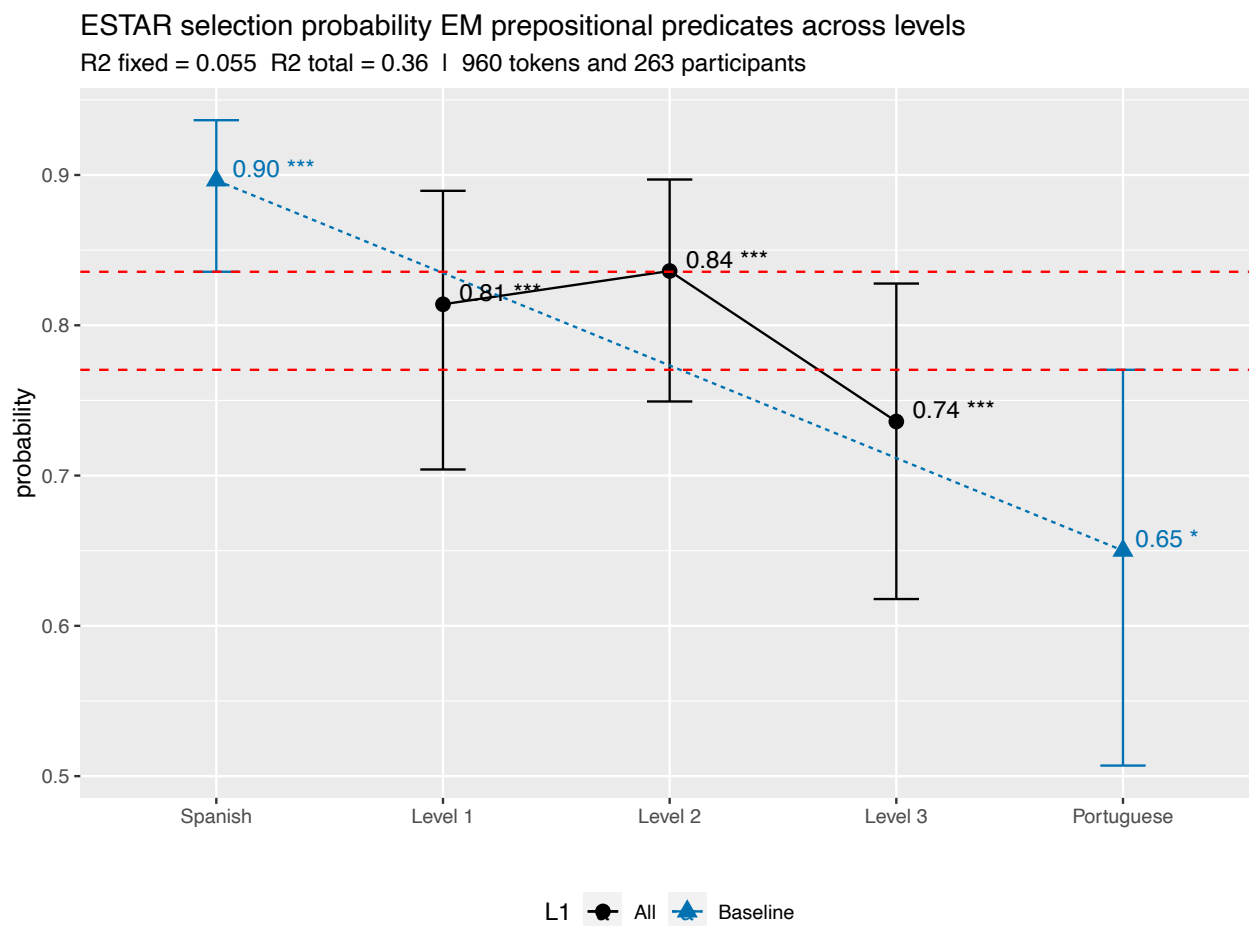


Figure 8.6: Logistic regression ESTAR selection probability estimates EM prepositional predicates. The dotted red lines mark lower confidence for Arizonian Spanish and upper confidence for Brazilian Portuguese, for reference.

[0.83, 0.93] for L1 English and Baseline Spanish respectively). The 95% confidence intervals for the other two groups, i.e., L1 Spanish (mean = .71, 95% CIs [0.53, 0.84]) and Heritage Spanish (mean = .77, 95% CIs [0.67, 0.85]) speakers, overlap with the target language (mean = .65, 95% CIs [0.51, 0.77]).

Table 8.13: Logistic regression results for ESTAR vs. SER with EM prepositional predicates across L1 background and baseline corpora.

L1	Probability	CIs	Pr(> z)	
Portuguese	0.65	[0.51, 0.77]	0.04	*
L1 English	0.86	[0.77, 0.91]	0.00	***
Heritage Spanish	0.77	[0.67, 0.85]	0.00	***
L1 Spanish	0.71	[0.53, 0.84]	0.02	*
Spanish	0.89	[0.83, 0.93]	0.00	***

8.4 Conclusion

Tables 8.14 and 8.15 show summaries of logistic regression results for Southern Arizona Spanish, L3 Portuguese, and Brazilian Portuguese, displaying copula preference (i.e., *ser* or *estar*) by predicate type. The estimated probability for *estar* selection is shown between parentheses. Empty cells (-) mean results for that corpus and predicate were not significantly different from chance. Cells with NA mean the results for that model did not converge.

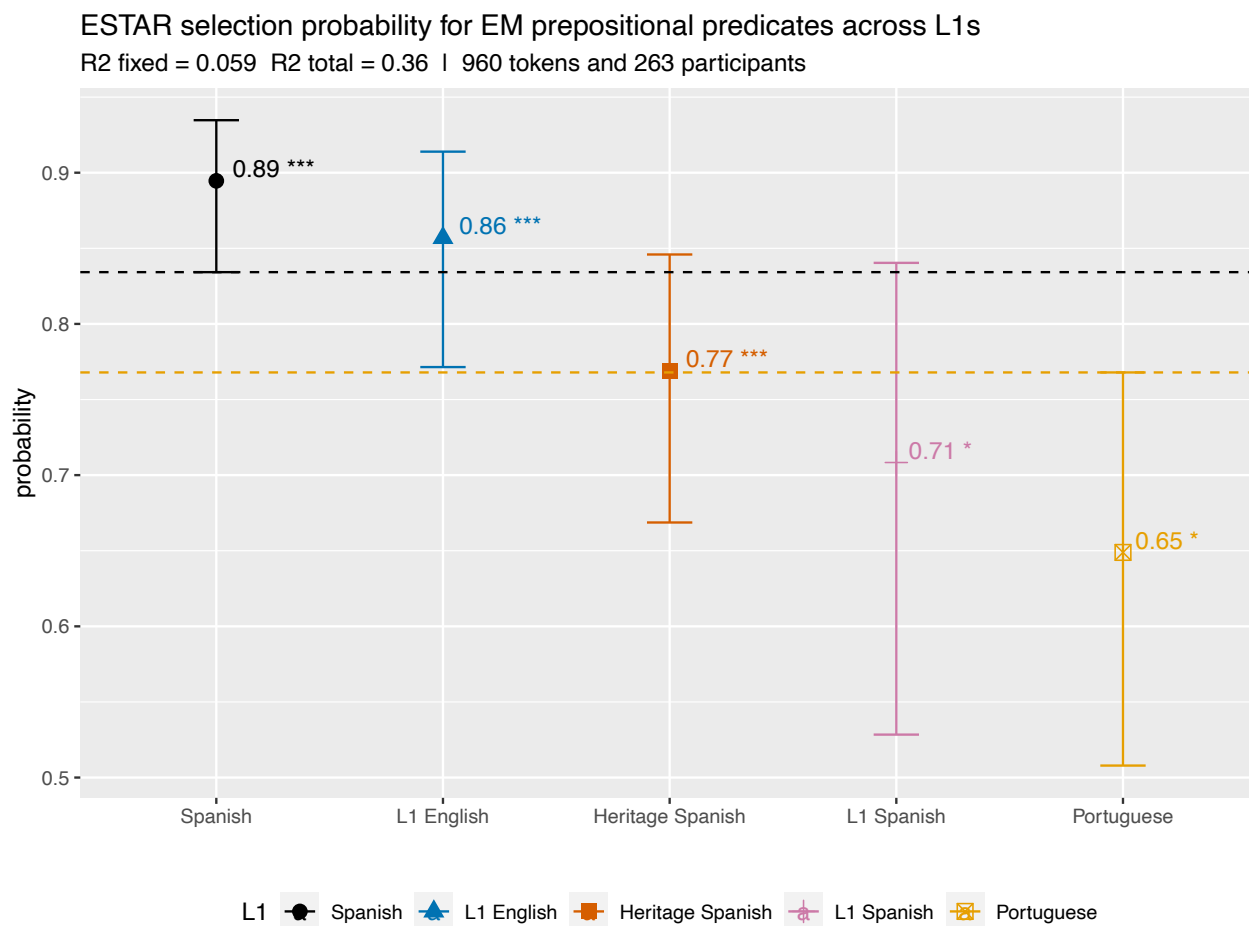


Figure 8.7: Logistic regression ESTAR selection probability estimates EM prepositional predicates. The dotted black line mark lower confidence for Arizonian Spanish and the dotted yellow line marks upper confidence for Brazilian Portuguese, for reference.

Table 8.14: Summary of logistic regression results across corpora.

Predicate Type	Arizona	Level 1	Level 2	Level 3	Brazil
adjectival: description	ser (0.12)	ser (0.05)	ser (0.05)	ser (0.15)	ser (0.26)
adjectival: evaluation	ser (0.05)	ser (0.01)	ser (0.01)	ser (0.00)	ser (0.31)
adjectival: verbal	estar (0.85)	-	-	estar (0.83)	estar (0.85)
prepositional: em	estar (0.90)	estar (0.81)	estar (0.84)	estar (0.74)	estar (0.65)

Table 8.15: Summary of logistic regression results across corpora.

Predicate Type	Arizona	L1 English	Spanish H.	L1 Spanish	Brazil
adjectival: description	ser (0.12)	ser (0.06)	ser (0.11)	ser (0.08)	ser (0.26)
adjectival: evaluation	ser (0.05)	NA	NA	NA	ser (0.31)
adjectival: verbal	estar (0.85)	estar (0.72)	-	estar (0.80)	estar (0.85)
prepositional: em	estar (0.90)	estar (0.86)	estar (0.77)	estar (0.71)	estar (0.65)

Results show that with description adjectives, verbal adjectives, and *em* prepositional predicates, L3 Portuguese production approximates the target language behavior across all three L3 Portuguese levels. That is not the case for adjectives of evaluation, with which L3 Portuguese learners show a strong preference for the default unmarked copula *ser* across all three levels instead of approximating the Portuguese baseline.

For *estar* selection with *em* prepositional predicates, L3 Portuguese production at the first two levels parallels the Spanish baseline corpus, indicating transfer from the use of the marked copula *estar* from Spanish. This transfer effect is stronger with L1 English (and thus L2 Spanish) speakers. Spanish transfer might also be present in the results for *estar* preference with adjectives of description, especially for the L1 Spanish and Heritage Spanish

speakers, who present overlapped *estar* probabilities with the Spanish baseline corpus. The L1 English speakers display lower probabilities of *estar* selection compared to the Spanish baseline, indicating the possibility of English influence instead. English might be also a source of influence for Heritage Spanish speakers when it comes to *estar* selection with verbal adjectives, since they display a much lower *estar* probability estimates than both Spanish and Portuguese baseline corpora.

The non-target-like preference for *estar* copula with evaluation adjectives might be an issue of salience. Since both *estar* and *ser* are used with adjectives of evaluation in Portuguese, the increased preference for *estar* in Portuguese (compared to Spanish) might not be salient enough for students to adjust their behavior over the course of the three L3 levels. The learners' overuse of *ser* might be evidence that they are acquiring *estar* from scratch, without Spanish transfer, and are thus replicating the stages of acquisition of L2 Spanish learners, who rely exclusively on *ser* in early stages of acquisition (Garavito & Valenzuela, 2006; Guntermann, 1992; Ryan & Lafford, 1992; VanPatten, 2010). In addition, since over-selection of *ser* with adjectives of evaluation is not ungrammatical, there is no reason for instructors to offer corrective feedback on its use. Contrast that with *em* prepositional predicates and verbal adjectives, which cause ungrammatical production if the wrong copula is chosen in certain contexts and are thus subjected to corrective feedback.

The results presented in this chapter provide evidence of hindering transfer from Spanish for L1 English speakers when selecting *estar* with *em* prepositional predicates and facilitative transfer from Spanish for L1 Spanish speakers when selecting *estar* with verbal adjectives. There is also evidence of transfer from Spanish for L1 Spanish and Heritage Spanish speakers with adjectives of description, and possible transfer for English for L1 English speakers.

CHAPTER 9

DISCUSSION

9.1 Introduction

This dissertation has investigated patterns of copula use across three languages. Both monolingual and bilingual native (L1) baseline corpora, in different modes, were used to compare copula use in Spanish, Portuguese, and English. With differences and similarities across these corpora established, I examined the development of Portuguese as a third language (i.e., L3 or any language acquired in adulthood after two or more languages have been previously acquired), and how L3 Portuguese production is affected by previously learned languages (i.e., Spanish and English). The baseline corpora were used to establish both language transfer and L3 development patterns.

The L3 Portuguese learner data analyzed consists of language that has been produced by students who fall under one of the following bilingual groups:

1. L1 English L2 Spanish (L1 English)
2. L1 Spanish L2 English (L1 Spanish)
3. L1 Spanish/English (i.e., heritage speakers of Spanish) (Heritage Spanish)

For transfer, the goal was to establish whether the source language was either an L1, an L2, or both (Bardel & Falk, 2007; De Angelis, 2007; Flynn et al., 2004; Rothman, 2014; Slabakova & Pilar García Mayo, 2017). The L1 Brazilian Portuguese baseline corpora was used to establish approximation to target language over proficiency levels (i.e., language development). As such, my research questions reflect the need to investigate initial transfer (i.e., cross-linguistic influence at lower levels of proficiency) separately from L3 development (i.e., differences in patterns of language use across proficiency levels). From a broad perspective, my two main research questions are divided into how patterns of copula use overlap and differ across 1)

the baseline corpora, and 2) L3 bilingual group and L3 proficiency level. The next sections in this chapter discuss each question separately. A final conclusion brings together main findings, research and pedagogical implications, and future directions.

9.2 Research Question #1

The first research question addresses copula patterns in language use across all three baseline languages:

1. How do the patterns of copula use differ across the target language (i.e., L1 Brazilian Portuguese), and the previously acquired languages (i.e., L1 Spanish and L1 English)?

Table 9.1 shows a summary of word embedding results for three Spanish (i.e., Southern Arizona, Mexico, and Spain) and the Portuguese (i.e., Brazil) baseline corpora, displaying copula preference (i.e., *ser* or *estar*) by adjective type. Cells with a - mean results for that corpus and adjective type were not significant, i.e., that corpus and adjective type combination display no preference for either copula. Cells with a X indicate that results for that combination of corpus and predicate type was not discussed in this dissertation.

Table 9.1: Summary of word embedding results for the Spanish corpora.

Predicate Type	Brazil	Arizona	Mexico	Spain
adjectival: age	-	ser	estar	-
adjectival: size	-	-	-	-
adjectival: evaluation	X	ser	-	ser
adjectival: physical appearance	estar	-	X	X
intensifier presence	ser	-	X	X
prepositional: em	-	estar	X	X

Table 9.2 shows a summary of logistic regression results for Southern Arizona Spanish and

Brazilian Portuguese, displaying copula preference (i.e., *ser* or *estar*) by predicate type. The estimated probability for *estar* selection is shown between parentheses.

Table 9.2: Summary of logistic regression results for baseline corpora. All results are significantly different from a chance probability.

Predicate Type	Brazil	Arizona
adjectival: description	ser (0.25)	ser (0.11)
adjectival: evaluation	ser (0.31)	ser (0.05)
adjectival: verbal	estar (0.85)	estar (0.85)
prepositional: em	estar (0.64)	estar (0.89)

I will focus my discussion on two pairwise comparisons: Spanish vs. Portuguese and English vs. Portuguese, organized by predicate type to address differences and similarities across corpora.

9.2.1 Adjectival Predicates

Logistic regression modeling results for copula as a binary choice as the dependent variable show that both baseline corpora (Spanish and Portuguese) display a preference for *ser* with all adjectival predicates aggregated, with a mean estimate probability of .31 for *estar* selection in Spanish (95% CIs [0.28, 0.35]) and .35 in Portuguese (95% CIs [0.30, 0.39]). In other words, approximately a third of the instances of *copula + adjectival predicates* constructions are realized with *estar*. However, when adjectival predicates are analyzed by adjective category (e.g., physical appearance, verbal), copula selection preference varies.

Results of *estar* embedding preference with different types of adjective embeddings include adjectives that have been shown to display different levels of preference for *estar* across different varieties of Spanish, including adjectives of age, size, and physical appearance (Bessett, 2015;

Cortés-Torres, 2004; Salazar, 2007; Silva-Corvalán, 1986). Adjective embeddings of age and size show no significant differences between the Southern Arizona and the Brazil corpora, due to high variability in the Brazil data. However, for these two types of adjectives, the copula verb embeddings in the Southern Arizona Spanish corpus result in a preference for *ser*, with log odd estimates of -0.03 for age adjectives and -0.02 for size adjectives (negative numbers mean a preference for *ser*). With adjectives of physical appearance, the Brazil corpus does show a preference for *estar* (log odd estimate of .06), but here the Arizona corpus does not display a preference for either copula (95% CIs [-0.02, 0.04]).

Adjectival predicates for evaluation and description were analyzed for *estar* preference as a binary choice (*ser* vs. *estar*) through logistic regression modeling. These specific adjective classes have been chosen due to their high frequency in the baseline corpora, which allows the logistic regression models to converge (i.e., there is a minimum number of tokens required). For both description and evaluation adjectives, Brazilian Portuguese shows a stronger preference for *estar* than the Arizonian Spanish corpus (probability estimates of .25 and .31 for Portuguese, and .11 and .05 for Spanish). As mentioned, copula choice with these adjectives indicates speakers' intent and whether that quality is meant to be permanent or temporary (Woolsey, 2008). These results indicate that input in the target language contains non-zero levels of *estar* with evaluation and description adjectives.

The results above showing weak *estar* preference with age, size, physical appearance, evaluation, and description adjectives in both Spanish and Portuguese match what is known about *copula+adjective* constructions in these languages. Whenever there is a possibility of choosing either copula verb, which is the case with most adjectival predicates, there is a difference in meaning that is realized with that choice, expressing the speaker intent (Moura, 2016; Schmitt & Miller, 2007; Sibaldo, 2011; Woolsey, 2008). Differences in meaning can only be accounted through hand coding of each copula construction taking into account the entire context of the discourse, including prosody, which is not accounted for in this dissertation. The verb *ser* is considered unmarked or underspecified for aspect in both Spanish and Portuguese, while *estar* indicates a state (i.e., property P holds at a certain point t in time) (Finnemann, 1990; Garavito & Valenzuela, 2006; Schmitt, 1992; Schmitt et al., 2004), so *ser* is more frequently

used.

Results for copula choice with verbal adjectives are also included here due to the high occurrence of *estar* with this adjective type in both languages. Estimated probabilities for *estar* from logistic regression modeling with corpus as a predictor (i.e., Arizonian Spanish versus Brazilian Portuguese) show that both languages display similar *estar* preferences with verbal adjectives in copula predicate position (mean estimated probability of .85 for both languages). These results again match what is known of copula use with verbal adjectives, since the copula *estar* is used with an adjective derived from an accomplishment verb (e.g., abierto *open*, cansado *tired*, and confundido *confused*), with only a few exceptions that show *ser* being more often used with verbal adjectives (e.g., casado *married*) (Carvalho & Bagno, 2015; Garavito & Valenzuela, 2006).

Regarding adjectives in predicate position in English copula (i.e., *to be*), Portuguese cognate (i.e., orthographic form similarity) versus non-cognate adjective preference was investigated. Results for word embeddings show that *cognate* preference varies across corpora according to their mode (i.e., writing versus spoken) and size, i.e., whether the corpus represents a specific speech community or is broadly representative of American English. English adjectives that are Portuguese cognate are slightly preferred with copula *be* embeddings for the CORE corpus while non-cognates are preferred for the BangorTalk Miami corpus. The mode here (writing vs. speech) explains these differences, with written production favoring Latin-based words (i.e. Portuguese cognate) such as *intelligent* over *smart*. The opposite pattern is produced in the spoken production in the the BangorTalk Miami corpus. For the Cambridge corpus, there is too much variation in the cognate embeddings, and no differences are found across cognate versus non-cognate words. The Cambridge corpus is much larger than the other two corpora (i.e., CORE and BangorTalk Miami), and it has a greater number of speakers represented as well.

9.2.2 Prepositional Predicates

The results of linear regression with copula embeddings representing *estar* preference show that the Southern Arizona Spanish corpus displays a stronger preference for *estar* with *em* prepositional predicates than the Brazil corpus (mean log odd estimate of .29 for the Spanish corpus and -0.03 for the Portuguese corpus). There is much more variability in the Brazil corpus (95% CIs [-0.35, 0.29]) than the Southern Arizona Spanish corpus (95% CIs [0.23, 0.36]) indicating a more consistent use of copula choice with *em* prepositional predicates in the Southern Arizona corpus.

Estar preference with *em* prepositional predicates is also analyzed using logistic regression modeling with copula choice as a binary dependent variable. Results here also show the Spanish corpus displaying a stronger preference for *estar*, with much higher probability for this copula with this type of predicate (estimate probability of .89) compared to Brazilian Portuguese (.64).

This stronger preference for *estar* in Spanish than in Portuguese matches what is reported in the literature. Prepositional phrases in the predicate position of a copula verb never allow for interchangeable use of *ser* and *estar* (Sibaldo, 2011). In Portuguese we use *ser* with locatives that are stationary in Portuguese (Moura, 2016), which differs from Spanish, e.g. Spanish: *¿Dónde está el baño?*, Portuguese: *Onde é o banheiro?* (i.e, Where is the bathroom?), which results in a stronger preference for *estar* in Spanish.

9.2.3 Nominal Predicates

Logistic regression results with copula use as a binary dependent variable show that there is a low probability of *estar* selection with nominal predicates in both Spanish and Portuguese (.07 and .01 respectively). These very low probabilities of *estar* with nominal predicates are consistent with what I discussed in the literature review, with nominal predicates being prescriptively used with *ser* as the only verb choice due to this construction's permanent status (Moura, 2016; Schmitt & Miller, 2007). However, the results presented in this dissertation

are not at zero, showing that there are usage exceptions to the prescriptive rule of *ser* with nominal predicates. These include slang expressions as in *ela está gente boa* (i.e., she is a nice person) in Portuguese and *cuando estaba chamaca* (i.e., when I was a child) in Spanish.

9.3 Research Question #2

The second research question addresses cross-linguistic influence in terms of transfer and L3 development:

2. How do patterns of L3 Portuguese production by Spanish-English bilinguals compare to the source languages (i.e., Spanish and/or English) versus the target language (i.e., Portuguese) at each of the three levels of L3 proficiency?

Table 9.3 shows a summary of word embedding results for Southern Arizona Spanish, L3 Portuguese, and Brazilian Portuguese, displaying copula preference (i.e., *ser* or *estar*) by predicate type. Empty cells (-) mean results for that corpus and adjective type were not significant, i.e., that corpus and predicate type combination display no preference for either copula.

Table 9.3: Summary of word embedding results for base-line and L3 Portuguese corpora.

Predicate Type	Arizona	L3_Portuguese	Brazil
adjectival: age	ser	-	-
adjectival: size	ser	-	-
adjectival: physical appearance	-	-	estar
prepositional: em	estar	-	-

Tables 9.4 and 9.5 show summaries of logistic regression results for Southern Arizona Spanish, L3 Portuguese, and Brazilian Portuguese, displaying copula preference (i.e., *ser* or *estar*) by predicate type. The estimated probability for *estar* selection is shown between parentheses.

Empty cells (-) mean results for that corpus and predicate were not significantly different from chance. Cells with NA mean the results for that model did not converge.

Table 9.4: Summary of logistic regression results across corpora.

Predicate Type	Arizona	Level 1	Level 2	Level 3	Brazil
adjectival: description	ser (0.12)	ser (0.05)	ser (0.05)	ser (0.15)	ser (0.26)
adjectival: evaluation	ser (0.05)	ser (0.01)	ser (0.01)	ser (0.00)	ser (0.31)
adjectival: verbal	estar (0.85)	-	-	estar (0.83)	estar (0.85)
prepositional: em	estar (0.90)	estar (0.81)	estar (0.84)	estar (0.74)	estar (0.65)

Table 9.5: Summary of logistic regression results across corpora.

Predicate Type	Arizona	L1 English	Spanish H.	L1 Spanish	Brazil
adjectival: description	ser (0.12)	ser (0.06)	ser (0.11)	ser (0.08)	ser (0.26)
adjectival: evaluation	ser (0.05)	NA	NA	NA	ser (0.31)
adjectival: verbal	estar (0.85)	estar (0.72)	-	estar (0.80)	estar (0.85)
prepositional: em	estar (0.90)	estar (0.86)	estar (0.77)	estar (0.71)	estar (0.65)

This dissertation hypothesized two scenarios for initial stages of L3 Portuguese production by Spanish-English bilinguals regarding *ser* and *estar* use: L3 Portuguese learners do 1) not transfer their copula knowledge from Spanish and thus replicate the stages described above (i.e., overuse of *ser* and gradual introduction of *estar*) or 2) transfer their copula knowledge from Spanish displaying *estar* use at initial staged of L3 production. For English transfer, the predictions were that L2 Portuguese learners either rely on English cognate adjectives in copula position or do not.

The results for all three bilingual groups show evidence of L3 Portuguese development for all

three groups of Spanish-English bilinguals. However, transfer from Spanish and English into L3 Portuguese production is not the same across all of these groups, varying also in degree depending on the copula construction. As such, these results conflict with the *Typological Primacy Model*, which predicts that L3 acquisition in adulthood starts off from a wholesale transfer of the pre-acquired language system that is most typologically similar to the target language (Rothman, 2010). According to this model, the three types of Spanish-English bilinguals included in this dissertation would transfer their entire knowledge of Spanish copula constructions to their L3 Portuguese use. Instead, Spanish status (i.e., L1, L2, or heritage) seems to affect what is transferred from Spanish (i.e., property-by-property transfer).

This dissertation offers support and expands on *The Parasitic Model* (Ecke, 2015; Ecke & Hall, 2014; Hall et al., 2009). As seen in the literature review, this model proposes three stages for L3 lexical acquisition that includes three aspects of lexical production (i.e., form, syntactic frame, concept/meaning). The first two stages are related to early L3 word learning, where a form representation is first established to a cognate form in either an L1 or an L2 and later form-frame-concept connections are built and rebuilt based on L3 development (including results from English cognate analysis). In addition to form-frame-concept connections based on cognate words, this dissertation offers evidence of connections based on selective attention (Ellis, 2006b), which takes into account factors such as unreliable mapping of language features (i.e., contingency) (Ellis, 2006a). The results in this dissertation across all three bilingual groups also highlight the importance of factors that are accounted by the *Scalpel model*, which argues that during L3 acquisition, L1 or L2 options relevant to the constructions being learned are extracted “with a scalpel-like precision” based on similarities across languages and structural linguistic complexity (Slabakova, 2016, p. 653).

The rest of this section is divided into the three sub questions under this second research question.

9.3.1 Which language is the source of initial transfer for each Spanish-English bilingual group?

Logistic regression results for *estar* preference as a binary choice show that L3 Portuguese learners display a strong preference for *ser* with description adjectives at the first two L3 Portuguese levels (estimate probability of *estar* choice at .05), which indicates no transfer from Spanish for adjectives of description. Results are similar for adjectives of evaluation, with all three L3 Portuguese levels displaying probabilities of *estar* selection with this type of adjective close to zero (.01 for first two levels, and .00 for level 3).

These results are similar to previous research in L2 Spanish acquisition by L1 English speakers, which has shown that adult learners go through similar stages of copula acquisition in Spanish: first there is over-use of *ser* due to its unmarkedness, then *estar* is gradually introduced in different contexts (Ryan & Lafford, 1992; VanPatten, 1985, 1987). The use of both copula *ser* and *estar* with adjectival predicates showing a differentiation in meaning in Spanish is one of the last structures to be acquired by L2 learners (VanPatten, 2010). The L3 Portuguese learners' overuse of *ser* shown in this dissertation might be evidence that they are acquiring *estar* from scratch (no transfer from Spanish).

No indication of transfer from Spanish is present in the logistic regression results of *estar* preference for verbal adjectives either. Verbal adjectives are mostly used with *estar* in both Portuguese and Spanish, but at the first two levels of L3 Portuguese proficiency learners do not show a preference for either verb (estimate probabilities for *estar* selection is at chance). It is worthwhile to note again that *estar* is the marked copula (Finnemann, 1990; Garavito & Valenzuela, 2006; VanPatten, 2010), and over-reliance on the unmarked copula *ser* is expected at lower levels when transfer is not present (Woolsey, 2008) .

In contrast to the no-transfer evidence discussed above, results for *estar* preference with *em* prepositional predicates show L3 Portuguese production at the first two levels parallels the Spanish baseline corpus. As mentioned during the discussion for the first research question, Brazilian Portuguese shows a weaker preference for *estar* with *em* prepositional predicates compared to Arizonian Spanish. At the first L3 Portuguese level, L3 Portuguese learners

show a very strong preference for *estar* with this type of predicate (mean *estar* probability of .81, 95% CIs [0.70, 0.89]) which approximates the probabilities for Arizonian Spanish (mean = .90, 95% CIs [0.84, 0.94]). This indicates transfer from the use of the marked copula *estar* from Spanish, instead of defaulting to the unmarked *ser*.

Regarding the preference of *English cognate* adjectives, the cross-linguistic influence of English seems to be stronger for the L1 English group, which shows a stronger preference for *English cognate* words compared to the Spanish Heritage group. Across the three levels, both L1 English and Spanish Heritage groups show a decrease for *English cognate* words as level increases.

Overall, results indicate that the source of initial transfer might be English or Spanish, depending on the copula construction.

9.3.2 At what point (i.e., first, second or third semester) do L3 Portuguese patterns of copula use become most similar to L1 Portuguese (as opposed to most similar to L1 Spanish or L1 English)?

Logistic regression results for *estar* preference as a binary choice show that L3 Portuguese production does not approximate the target language for adjectives of evaluation, with L3 Portuguese learners showing a strong preference for the default unmarked copula *ser* across all three levels instead of approximating the Portuguese baseline. Results for the Brazilian Portuguese baseline indicate that input in the target language contains non-zero levels of *estar* with evaluation adjectives. As so, L3 Portuguese learners were expected to use *estar* with some of these adjectives at non-zero levels as well, at some point in their L3 development. But they never do. As mentioned, both *estar* and *ser* are used with this type of adjective in Portuguese. It is possible that the increased use of *estar* with evaluation adjectives in Portuguese is a difference from Spanish that is too subtle for L3 Portuguese learners to perceive (Geeslin & Guijarro-Fuentes, 2006). In addition, there is no reason for instructors to offer corrective feedback on *ser* overuse with evaluation adjectives, since the use of *ser* with adjectives of evaluation is never ungrammatical.

Results for *estar* preference in logistic regression modeling with copula choice as a binary dependent variable show that with description adjectives, verbal adjectives, and *em* prepositional predicates, L3 Portuguese production approximates the target language behavior at the third L3 Portuguese level. For all of these three constructions, no differences are present between the first two L3 Portuguese levels. Greater gains (i.e., difference between first and last L3 Portuguese levels) are seen on *estar* use with verbal adjective (.25 gain toward target language pattern) than with *em* prepositional predicates (.05 gain) and description adjectives (.10 gain). One possible explanation for this is that copula use with verbal adjectives can cause ungrammatical production if the wrong copula is chosen in certain contexts and are thus subjected to corrective feedback. Although this is also the case with *em* prepositional predicates, this type of error (i.e., using *estar* with stationary referents in locative constructions) might not be as frequent or salient as *ser* with adjectives derived from accomplishment verbs.

Whether or not instructors provide negative evidence for ungrammatical use of *estar* with *em* prepositional predicates, L3 Portuguese learners move away from Spanish patterns towards Portuguese patterns of copula use. As seen in the literature review, *em* prepositional locatives are always predicate to *estar* copula in Spanish while there is an extra factor to be considered in Portuguese (i.e., whether the referent is stationary) which results in the selection of *ser* instead (Sibaldo, 2011). As a result, for locatives, there is a conflict of mappings between the source and the target languages. Going back to *The Parasitic Model*, in the first stage L3 Portuguese learner recognize the Portuguese frame *estar em* as cognate with *estar en* in Spanish, and transfer this construction to all locatives in their L3 Portuguese at first, without accounting for the type of referent (stationary vs. mobile). At later stages, L3 learners reanalyze this connection to make the distinction between *ser* and *estar* according to the referent.

Taking all results into consideration, L3 Portuguese patterns of copula use become most similar to L1 Portuguese at the third L3 Portuguese level, with little difference in development between level 1 and level 2. These results confirm the need to look at different levels of proficiency (Alonso & Rothman, 2016; Green, 2017; Slabakova, 2016; Slabakova & Pilar

García Mayo, 2017), from beginners (i.e., first semester Portuguese for the purposes of this dissertation) to more advanced learners (i.e., third semester Portuguese). Transfer effects are stronger at Level 1, which shows evidence of initial transfer, but some of these effects remain from some groups and some copula constructions throughout L3 development (Slabakova & Pilar García Mayo, 2017).

9.3.3 How similar are the L3 development paths across the three Spanish-English bilingual groups?

L3 Portuguese development paths across the three Spanish-English bilingual groups is at times similar and at times different depending on the copula structure. Logistic regression results for *estar* preference as a binary choice show that the L1 Spanish and Heritage Spanish speakers display overlapping *estar* probabilities with the Spanish baseline corpus with adjectives of description. The L1 English speakers display lower probabilities of *estar* selection with adjectives of description compared to the Spanish baseline, indicating the possibility of English interference instead.

English might be also a source of interference for Heritage Spanish speakers when it comes to *estar* selection with verbal adjectives, since they display much lower *estar* probability estimates than both Spanish and Portuguese baseline corpora, while probabilities for both L1 English and L1 Spanish speakers overlap partially with the baseline corpora. What is interesting here is that cognate status with Spanish should help learners guess the right copula to use with each verbal adjective (i.e., Portuguese and Spanish constructions overlap), but somehow learners fail to make that form-frame-concept connection. This is especially true for Heritage Spanish speakers, which present lower probabilities of *estar* selection with verbal adjectives (95% CIs [0.45, 0.66]) compared to both L1 English and L1 Spanish speakers (95% CIs of [0.61, 0.81] and [0.64, 0.90] respectively).

Results for *estar* preference in logistic regression modeling with copula choice as a binary dependent variable show that with *em* prepositional predicates the transfer effect from Spanish to L3 Portuguese production is stronger with L1 English speakers. The fact that L1 English

speakers (and thus L2 Spanish speakers) stick the most to the patterns of Spanish compared to L1 Spanish and Heritage Spanish speakers indicate that *em* locatives are harder for L1 English speakers to reanalyze in their L3 Portuguese development. L1 Spanish speakers, on the other hand, have overlapping probabilities of *estar* selection with the Portuguese baseline for this copula construction. The probabilities of *estar* selection for Heritage Spanish speakers for this copula structure sit in-between the Spanish and Portuguese baseline corpora.

9.4 Limitations

The main limitation of this dissertation is the exclusive use of untimed written (i.e., offline) L3 Portuguese production, without the inclusion of L3 online or spoken language. Since the baseline corpora in Spanish and in Portuguese are oral, another limitation of this research is the comparison of spoken L1 corpora to written L2 corpora. In addition, the Spanish and English baselines used as baseline for this research are not produced by the same participants as the Spanish-English bilinguals in the L3 Portuguese corpus.

The quantitative analysis in this dissertation did not take into account issues of pragmatics, including speaker intent when there is a choice between *ser* and *estar* with certain adjectives in Portuguese. While this research project assumes that learners had the chance to use *estar* based on the varied topics and genres represented in the L3 Portuguese learner corpus, actual differences in meaning realized through copula choice could only have been accounted through hand coding of each construction, which is a time-consuming process that requires labeling a number of discourse features. These factors are not considered in this dissertation. Nevertheless, coding each instance of copula for topic and intent would result in a better picture for why learners are not choosing *estar* at the same level as the target language when that choice is available to them (i.e., with adjectival predicates of evaluation).

This dissertation does not address the language development of individuals, as suggested by Larsen-Freeman (2015), and only presents results in aggregated manner. Results do show a lot of individual variability in the L3 learner data, and I believe that a more qualitative analysis of a few participants would add to the understanding on both cross-linguistic influences and

development in L3 Portuguese acquisition.

9.5 Implications

The results discussed here have research and pedagogical implications. For pedagogical implications, there is a need for explicit instruction of Portuguese copula use with both verbal adjectives, which seems to not be facilitated by Spanish transfer, and *em* locatives, which shows hindering Spanish transfer effects into Portuguese development. For research implications, there is a need for linguistic analysis that allow for statistical comparison of language use across groups.

I agree with Slabakova (2016) who calls for more flexible L3 acquisition models that take a variety of factors into account, not only language status (i.e., L1 versus L2) and typological similarities across languages. These factors include linguistic form difficulty (complexity of form, meaning, and form-meaning relationship) and cognitive prominence, e.g. age of onset (Slabakova, 2012, 2016). These multiple factors can only be accounted by more sophisticated methods of language analysis, including the two methods used in this dissertation (i.e., word embeddings and logistic regression). Linear regression combined with word embeddings is a powerful way to analyze main differences across large corpora. Logistic regression is more labor intensive, because it requires manual coding of some contextual features, but it allows for a more detailed account of differences across groups. The two methods of analysis combined provide different but complementary perspectives on language use.

As seen in the discussion of results, this dissertation offers support to L3 acquisition models that take into consideration structural characteristics of individual constructions, and how similar or different these are between source and target languages. Only property-by-property models such as *The Parasitic Model* (Ecke, 2015; Ecke & Hall, 2014; Hall et al., 2009) and the *Scalpel Model* (Slabakova, 2016) grant room for these factors. The results presented here also provide support for the consideration of other factors like construction frequency and unambiguous input (Ellis, 2006b; Slabakova, 2016).

Probabilistic language processing (Ellis, 2006b, 2006a) might also play a role in the non-source-like and non-target-like *estar* copula selection with verbal adjectives. For example, the overall probability of *estar* selection with adjectives in general in both Spanish and Portuguese is low (.31 for Spanish and .35 for Portuguese). However, for verbal adjectives *estar* selection is high in both languages (.85). At lower levels of proficiency, L3 Portuguese learners do not commit to a specific copula verb with verbal adjectives (95% CIs of [0.43, 0.65] for level 1 and [0.47, 0.68] for level 2), which points to the learners' defaulting to *ser* to match the overall low frequency of *estar* with adjectives. At level 3, however, L3 Portuguese learners approximate the higher frequencies of *estar* selection found in the target language (95% CIs of [0.74, 0.89]) while keeping the frequency of *estar* selection low with other adjectives such as evaluation. The implication here includes pedagogical intervention comprising explicit instruction on the meaning restriction of individual verbal adjectives to establish the distinction between adjectives such as *avanzada* (advanced) in constructions like in *a aula é avanzada* (the class is advanced) and *assustado* (scared) like in *estou assustado* (I am scared).

Pedagogical implications for L3 Portuguese instruction also include the need for making the difference between locatives in Portuguese explicit, and calling students' attention to the fact that while we say *meu amigo está ali na esquina* (i.e., my friend is around the corner) when the referent is mobile like a person, we say *meu apartamento é ali na esquina* (i.e., my apartment is around the corner) when the referent is stationary.

This dissertation used two different methods of linguistic analysis that allow for statistical comparison of language use across groups, 1) word embeddings, and 2) logistic regression. Triangulating results with the two methods provides a better picture of the overall data. Logistic regression is a more attested method of language analysis, and by using it side-by-side with word embeddings, it was possible to attest the results obtained with word embeddings. That was clearly seen with *em* prepositional predicates, in which the word embeddings showed that the Southern Arizona Spanish corpus showed a clear preference for *esta* with this type of predicate while there was no preference for either copula with *em* prepositional predicates in the Brazilian Portuguese corpus. The logistic regression results mirrored the word embeddings, showing that Southern Arizona Spanish displayed a strong preference for

estar with *em* prepositional predicates (at an estimated probability of 89% for *estar* selection) while the Brazilian Portuguese displayed a much weaker preference for *estar* with the same predicate type (at an estimated probability of 64%). These results, from both methods, in turn match what we know about copula predicates in Spanish versus Portuguese, with Portuguese making use of both *ser* and *estar* with locatives and Spanish using only *estar* with locatives (Moura, 2016; Sibaldo, 2011). This difference implies an overall preference for *estar* copula with *em* prepositionals in Spanish (VanPatten, 2010).

Word embeddings require larger corpora, while logistic regression requires some manual checking of individual constructions, which is labor intensive. However, logistic regression results, at a smaller scale, can validate word embedding results at a much larger scale, as discussed above. Word embeddings also present an opportunity for the investigation of individual lexical items, which was beyond the scope of this dissertation, but an avenue to be pursued in the future.

9.6 Future Research

Preliminary analysis of individual participants' coefficients (not discussed in this dissertation) from the resulting logistic regression models shows a bi-modal distribution across all three types of Spanish-English bilinguals, which divides the L3 Portuguese learners into those who at the third L3 Portuguese level approximate the Portuguese copula use baseline and those who are still struggling to match the baseline. My most immediate future research include the implementation of a qualitative analysis of a few of longitudinal students in each of these groups to shed light on individual differences.

In addition, I agree with Alonso & Rothman (2016), who argue that more data from larger groups of L3 learners and from a larger range of morphosyntactic properties is needed to build a complete theory of L3 acquisition. My plan is to replicate this research on copula use on L3 Portuguese spoken data, once that data has been acquired and transcribed. In addition, I believe that the study of the L3 acquisition of other language features (e.g., obligatory contraction in Portuguese, nominal agreement) can benefit from the two methods

of quantitative analysis that I employed in this dissertation (i.e., word embeddings and logistic regression).

Finally, with the growth of MACAWS (Multilingual Academic Corpus of Assignments - Writing and Speech) (Staples et al., 2019), which the L3 learner Portuguese corpus used in this dissertation is now a part of, I plan to expand my L3 acquisition research to include other L3 languages learned by Spanish-English bilinguals. Also, the issue of too much variability in the learner data will attenuate with more data available in L3 Portuguese in MACAWS. It will then be possible to employ multivariate regression methods, and include in the analysis more linguistic contextual factors (e.g., subject type of each copula construction). Following the study on transfer of variable grammars in third language acquisition by Ortin & Fernandez-Florez (2019), I plan to compare the variable grammars of copula constructions across the same corpora used in this dissertation.

APPENDIX A

Adjective Restrictions in Spanish

A.1 Adjectives used with *ser* only

Spanish Word	Translation
apto	<i>suitable</i>
auténtico	<i>authentic</i>
búlgaro	<i>Bulgarian</i>
cauto	<i>cautious</i>
constante	<i>constant, persevering</i>
cuidadoso	<i>careful</i>
culpable	<i>guilty</i>
(des)cortés	<i>(im)polite</i>
(des)leal	<i>(dis)loyal</i>
español	<i>Spanish</i>
evidente	<i>evident</i>
falso	<i>false, forged</i>
fiel	<i>faithful</i>
(in)prudente	<i>(im)prudent</i>
(in)discreto	<i>(in)discreet</i>
(in)capaz	<i>(un)able, (in)capable</i>
(in)justo	<i>(un)fair</i>
(in)moral	<i>(im)moral</i>
(in)mortal	<i>(im)mortal, eternal</i>
inocente	<i>innocent</i>
inteligente	<i>intelligent</i>

Spanish Word	Translation
(in)necesario	<i>(un)necessary</i>
presumido	<i>arrogant, vain</i>
semanal	<i>weekly</i>
socialista	<i>Socialist</i>

A.2 Adjectives used with *estar* only

Spanish Word	Translation
absorto	<i>absorbed, captivated</i>
angustiado	<i>worried, distressed</i>
asombrado	<i>astonished</i>
ausente	<i>absent; distracted</i>
contento	<i>happy</i>
desnudo	<i>naked</i>
descalzo	<i>barefoot</i>
enfermo	<i>ill</i>
enojado	<i>angry</i>
harto	<i>fed up</i>
lleno	<i>full</i>
maltrecho	<i>battered</i>
muerto	<i>dead</i>
perplejo	<i>perplexed</i>
presente	<i>present</i>
quieto	<i>still</i>
satisfecho	<i>satisfied</i>
solo	<i>alone</i>

REFERENCES

- Alonso, J. G., & Rothman, J. (2016). Coming of age in l3 initial stages transfer models: Deriving developmental predictions and looking towards the future. *International Journal of Bilingualism*, 21(6), 683–697.
- Bardel, C., & Falk, Y. (2007). The role of the second language in third language acquisition: The case of germanic syntax. *Second Language Research*, 23(4), 459–484.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823*.
- Bessett, R. M. (2015). The extension of estar across the mexico-us border: Evidence against contact-induced acceleration. *Sociolinguistic Studies*, 9(4), 421.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing.
- Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora*, 10(1), 11–45.
- Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. *Corpus Linguistics: An International Handbook*, 2, 1286–1304.
- Bick, E. (2000). *The parsing system palavras - automatic grammatical analysis of portuguese in a constraint grammar framework*. Aarhus, DK: Aarhus University Press.
- Bohnacker, U. (2006). When swedes begin to learn german: From v2 to v2. *Second Language Research*, 22(4), 443–486. Retrieved from <http://ezproxy.library.arizona.edu/login?url=https://search.proquest.com/docview/200203197?accountid=8360>
- Bolker, B. (2018). Formulas and contrasts for linear models. *Ecology and Evolution*, 1, 103–113.
- Bybee, J., & Eddington, D. (2006). A usage-based approach to spanish verbs of 'becoming'.

Language, 82(2), 323–355.

Cabrelli Amaro, J. (2017). Testing the phonological permeability hypothesis: L3 phonological effects on L1 versus L2 systems. *International Journal of Bilingualism*, 21(6), 698–717.

Cabrelli Amaro, J., Amaro, J. F., & Rothman, J. (2015). The relationship between L3 transfer and structural similarity across development: Raising across an experimenter in Brazilian Portuguese. In H. Peukert (Ed.), *Transfer effects in multilingual language development* (Vol. 4, pp. 21–52). John Benjamins Publishing Company.

Cambridge. (2004). The Cambridge-Cornell corpus of spoken North American English. Retrieved from https://www.cambridge.org/elt/corpus/corpora_spoken_north_am.htm

Carvalho, A. (2012). Corpus del español en el sur de Arizona (CESA). Retrieved from cesa.arizona.edu

Carvalho, A. M., & Bacelar Da Silva, A. J. (2006). Cross-linguistic influence in third language acquisition: The case of Spanish-English bilinguals' acquisition of Portuguese. *Foreign Language Annals*, 39(2), 185–202.

Carvalho, O. L. S., & Bagno, M. (2015). Gramática brasileira para hablantes de español. *São Paulo: Parábola*.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 740–750). Doha, Qatar: Association for Computational Linguistics.

Child, M. W. (2014). *Cross-linguistic influence in L3 Portuguese acquisition: Language learning perceptions and the knowledge and transfer of mood distinctions by three groups of English-Spanish bilinguals* (PhD thesis). University of Arizona.

Cortés-Torres, M. (2004). ¿Ser o estar? La variación lingüística y social de estar más adjetivo en el español de Cuernavaca, México. *Hispania*, 788–795.

De Angelis, G. (2007). *Third or additional language acquisition*. Multilingual Matters.

Deuchar, M. (2011). The Miami corpus of Spanish-English bilingual speech.

- Ecke, P. (2015). Parasitic vocabulary acquisition, cross-linguistic influence, and lexical retrieval in multilinguals. *Bilingualism: Language and Cognition*, 18(2), 145–162.
- Ecke, P., & Hall, C. J. (2014). The parasitic model of L2 and L3 vocabulary acquisition: Evidence from naturalistic and experimental studies. *Fórum Lingüístico*, 11(3), 360–372.
- Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1), 1–24.
- Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, saliency, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, 27(2), 164–194.
- Falk, J. (1979). Visión de norma general versus visión de norma individual: Ensayo de explicación de la oposición ser/estar en unión con adjetivos que denotan belleza y corpulencia. *Studia Neophilologica*, 51(2), 275–293.
- Falk, Y., & Bardel, C. (2010). The study of the role of the background languages in third language acquisition: The state of the art. *International Review of Applied Linguistics in Language Teaching*, 48(2-3), 185–219.
- Fernández, F. M. (1996). Metodología del "proyecto para el estudio sociolingüístico del español de España y de América". *Lingüística*, (8), 257–287.
- Finnemann, M. D. (1990). Markedness and learner strategy: Form-and meaning-oriented learners in the foreign language context. *The Modern Language Journal*, 74(2), 176–187.
- Flynn, S., Foley, C., & Vinnitskaya, I. (2004). The cumulative-enhancement model for language acquisition: Comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses. *International Journal of Multilingualism*, 1(1), 3–16.
- Forsyth, H. (2014). The influence of L2 transfer on L3 English written production in a bilingual German/Italian population: A study of syntactic errors. *Open Journal of Modern Linguistics*, 4(3), 429–456.

- Fox, J., & Monette, G. (2002). *An r and s-plus companion to applied regression*. Sage.
- Garavito, J. B. de, & Valenzuela, E. (2006). The status of ser and estar in late and early bilingual l2 spanish. In C. A. Klee & T. L. Face (Eds.), *Selected proceedings of the 7th conference on the acquisition of spanish and portuguese as first and second languages* (pp. 100–109). Somerville, MA: Cascadilla Proceedings Project.
- García Mayo, M. del P., & Slabakova, R. (2015). Object drop in l3 acquisition. *International Journal of Bilingualism*, *19*(5), 483–498.
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.
- Gauthier, J. (2016). Spanish faq for stanford corenlp, parser, pos tagger, and ner. London: Longman. Retrieved from <https://nlp.stanford.edu/software/spanish-faq.html>
- Geeslin, K. L., & Guijarro-Fuentes, P. (2006). Second language acquisition of variable structures in spanish by portuguese speakers. *Language Learning*, *56*(1), 53–107.
- Geeslin, K. L., & Guijarro-Fuentes, P. (2008). Variation in contemporary spanish: Linguistic predictors of estar in four cases of language contact. *Bilingualism: Language and Cognition*, *11*(3), 365–380.
- Giancaspro, D., Halloran, B., & Iverson, M. (2015). Transfer at the initial stages of l3 brazilian portuguese: A look at three groups of english/spanish bilinguals. *Bilingualism: Language and Cognition*, *18*(2), 191–207.
- Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *Computation and Language*, *1*, 1–5.
- Green, D. W. (2017). Trajectories to third-language proficiency. *International Journal of Bilingualism*, *21*(6), 718–733.
- Gries, S. T., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9*(1), 109–136.

- Guntermann, G. (1992). An analysis of interlanguage development over time: Part ii, ser and estar. *Hispania*, 75(5), 1294–1303.
- Hall, C. J., Newbrand, D., Ecke, P., Sperr, U., Marchand, V., & Hayes, L. (2009). Learners' implicit assumptions about syntactic frames in new l3 words: The role of cognates, typological proximity, and l2 status. *Language Learning*, 59(1), 153–202.
- Håkansson, G., Pienemann, M., & Sayehli, S. (2002). Transfer and typological proximity in the context of second language processing. *Second Language Research*, 18(3), 250–273. Retrieved from <http://ezproxy.library.arizona.edu/login?url=https://search.proquest.com/docview/200245255?accountid=8360>
- Ionin, T., Montrul, S., & Santos, H. (2011). Transfer in l2 and l3 acquisition of generic interpretation. In N. D. K. Mesh & H. Sung (Eds.), *Boston university conference on language development* (pp. 283–295). Cascadilla Press.
- Iverson, M. (2009). Competing sla hypotheses assessed: Comparing heritage and successive spanish bilinguals of l3 brazilian portuguese. *Minimalist Inquiries into Child and Adult Language Acquisition*, 35.
- Larsen-Freeman, D. (2015). Saying what we mean: Making a case for 'language acquisition' to become 'language development'. *Language Teaching*, 48(4), 491–505.
- Levy, O., & Goldberg, Y. (2014). Dependencybased word embeddings. In *In acl*.
- Looney, D., & Lusin, N. (2019). Enrollments in languages other than english in united states institutions of higher education, summer 2016 and fall 2016. In *Modern language association*. ERIC.
- Lucas, C., & Tingley, D. (2014). *TranslateR: Bindings for the google and microsoft translation apis*. Retrieved from <https://CRAN.R-project.org/package=translateR>
- MacWhinney, B. (2014). *The childe's project: Tools for analyzing talk, volume i: Transcription format and programs*. London, UK: Psychology Press.
- Maimone, L. L. (2017). *The role of crosslinguistic influence from l2 spanish, type of linguistic*

- item, and aptitude in the learning stages of l3 portuguese forms: An exploratory study.* (PhD thesis). Georgetown University, Washinton, DC.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, *abs/1310.4546*. Retrieved from <http://arxiv.org/abs/1310.4546>
- Montrul, S., Dias, R., & Santos, H. (2010). Clitics and object expression in the l3 acquisition of brazilian portuguese: Structural similarity matters for transfer. *Second Language Research*, *27*(1), 21–58.
- Moro, A. (2000). *Dynamic antisymmetry* (Vol. 38). MIT Press.
- Moura, D. (2016). A predicação copulativa em português brasileiro e em espanhol. *Revista Do GELNE*, *9*(1/2), 67–76.
- Norris, J. M. (2015). Discriminant analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (305th-328th ed.). Routledge.
- Ortin, R., & Fernandez-Florez, C. (2019). Transfer of variable grammars in third language acquisition. *International Journal of Multilingualism*, *16*(4), 442–458.
- PRESEEA. (2014). Corpus del proyecto para el estudio sociolingüístico del español de españa y de américa. Retrieved from <http://preseea.linguas.net>
- Ramirez-Gelpi, A. S. (1997). *The acquisition of ser and estar among adult native english speakers learning spanish as a second language.* (PhD thesis). University of Southern California.
- Raso, T., & Mello, H. (2012). *C-oral-brasil: Corpus de referência do português brasileiro falado informal* (First). Belo Horizonte, MG: Editora UFMG.
- Rebelo, I., & Osório, P. (2006). Usos do verbo " ficar " no português do brasil: Classificação e análise. *Gragoatá*, *11*(21).
- Rebouças, R. A. F. (2018). Ser, estar e ficar em construções com adjetivos e particípios. *ElingUP: Revista Eletrônica de Linguística Dos Estudantes Da Universidade Do Porto*, *6*.

- Ribeiro, R. M. P. (2004). A expansão de sentidos do verbo ficar e os mecanismos responsáveis pela organização cognitiva de suas significações. *Revista Eletrônica Do Instituto de Humanidades*, 2(8), 1–8.
- Rothman, J. (2010). L3 syntactic transfer selectivity and typological determinacy: The typological primacy model. *Second Language Research*, 27(1), 107–127.
- Rothman, J. (2014). Linguistic and cognitive motivations for the typological primacy model (tpm) of third language (l3) transfer: Timing of acquisition and proficiency considered. *Bilingualism: Language and Cognition*, 18(2), 179–190.
- Ryan, J. M., & Lafford, B. A. (1992). Acquisition of lexical meaning in a study abroad environment: Ser and estar and the granada experience. *Hispania*, 75(3), 714–722.
- Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.
- Salazar, M. L. (2007). Esta muy diferente a como era antes ser and estar. *Spanish in Contact: Policy, Social and Linguistic Inquiries*, 22, 345.
- Schmid, H. (2019). Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd international conference on digital access to textual cultural heritage* (pp. 133–137). ACM.
- Schmitt, C. (1992). Ser and estar: A matter of aspect. In *Proceedings of nels* (Vol. 22, pp. 411–26).
- Schmitt, C. (2013). When stay and become are the same verb: The case of ficar. *ZAS Papers in Linguistics*, 227–255.
- Schmitt, C., Holtheuer, C., & Miller, K. (2004). Acquisition of copulas ser and estar in spanish: Learning lexico-semantics, syntax and discourse. In *Proceedings of boston university conference on language development*. Cascadilla press, somerville, ma.
- Schmitt, C., & Miller, K. (2007). Making discourse-dependent decisions: The case of the

- copulas ser and estar in spanish. *Lingua*, 117(11), 1907–1929.
- Schuster, S., & Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the tenth international conference on language resources and evaluation (lrec'16)* (pp. 2371–2378).
- Schwartz, B. D., & Sprouse, R. A. (1994). Word order and nominative case in nonnative language acquisition: A longitudinal study of (l1 turkish) german interlanguage. *Language Acquisition Studies in Generative Grammar*, 31(4), 71–89.
- Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access model. *Second Language Research*, 12(1), 40–72.
- Sibaldo, M. A. (2011). Para uma sintaxe diacrônica das sentenças copulares do português. *Revista Leitura*, 1(47).
- Silva-Corvalán, C. (1986). Bilingualism and language change: The extension of estar in los angeles spanish. *Language*, 62(3), 587–608.
- Slabakova, R. (2012). L3/Ln acquisition: A view from the outside. In J. C. Amaro, S. Flynn, & J. Rothman (Eds.), *Third language acquisition in adulthood* (Vol. 46, pp. 115–140). John Benjamins Publishing Company.
- Slabakova, R. (2016). The scalpel model of third language acquisition. *International Journal of Bilingualism*, 21(6), 651–665.
- Slabakova, R., & García Mayo, M. del P. (2015). The l3 syntax–discourse interface. *Bilingualism: Language and Cognition*, 18(2), 208–226.
- Slabakova, R., & Pilar García Mayo, M. del. (2017). Testing the current models of third language acquisition. In T. Angelovska & A. Hahn (Eds.), *L3 syntactic transfer: Models, new developments and implications*. John Benjamins Publishing Company.
- Soschen, A. (2002). *On subjects and predicates in russian: Syntax/semantics interface* (PhD thesis). University of Ottawa; Universal-Publishers, University of Ottawa.

- Staples, S., Novikov, A., Picoral, A., & Sommer-Farias, B. (2019). Multilingual academic corpus of assignments - writing and speech. Retrieved from <https://macaws.corporaproject.org/>
- Staples, S., & Reppen, R. (2016). Understanding first-year l2 writing: A lexico-grammatical analysis across l1s, genres, and language ratings. *Journal of Second Language Writing, 32*, 17–35.
- Tagliamonte, S. (2012). *Sociolinguistics as language variation and change*. Wiley-Blackwell.
- Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics, 11*(1/2), 61–82.
- VanPatten, B. (1985). Acquisition of ser and estar by adult learners of spanish: A preliminary investigation of transitional stages of competence. *Hispania, 68*(2), 399–406.
- VanPatten, B. (1987). Classroom learners' acquisition of ser and estar: Accounting for developmental patterns. In B. VanPatten, T. R. Dvorak, & J. F. Lee (Eds.), *Foreign language learning* (pp. 19–32). Newbury: Rowley.
- VanPatten, B. (2010). Some verbs are more perfect than others: Why learners have difficulty with ser and estar and what it means for instruction. *Hispania, 93*(1), 29–38.
- Wilson, D. (2010). *Formulaic language and adjective categories in eight centuries of the spanish expression of 'becoming'/quedar (se)/+ adj* (PhD thesis). University of New Mexico.
- Wilson, D. V. (2014). *Categorization and constructional change in spanish expressions of 'becoming'*. Brill.
- Winter, B. (2019). *Statistics for linguists: An introduction using r*. Routledge.
- Woolsey, D. (2008). From theory to research: Contextual predictors of “estar+ adjective” and the study of the sla of spanish copula choice. *Bilingualism: Language and Cognition, 11*(3), 277–295.
- Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for

phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1393–1398).