The development of sC onset clusters in Spanish English

Claudia Ivette Escartín Ortiz

A Thesis

in

The Department

of

Education

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Arts (Applied Linguistics) at
Concordia University
Montreal, Quebec, Canada

July 2005

Your file Votre référence
ISBN: 978-0-494-16241-5
Our file Notre référence
ISBN: 978-0-494-16241-5

Abstract

The development of sC onset clusters in Spanish English

Claudia Ivette Escartín Ortiz

This thesis investigates the variable phonology that characterizes the grammar of learner speech (interlanguage) in the context of data produced by native speakers of Spanish in a classroom environment. Specifically, the study follows the development of English /s/ plus consonant clusters (sC henceforth) across three levels of proficiency: beginners, intermediate and advanced. Because these clusters are absent in Spanish, learners syllabify these constituents with a preceding epenthetic [e] ([e]-epenthesis); as seen in the word 'snake' /snejk/ → [es.nejk]. The production of sC onset clusters is a variable interlanguage phenomenon influenced not only by the L1 interference, as implied above, but also by linguistic and extralinguistic factors. The results of the GoldVarb 2001 statistical analysis indicate that sC clusters occur more frequently in the following contexts: (1) when the preceding segment is a vowel; (2) in /s/ plus nasal sequences; (3) in more formal stylistic environments; and (4) in higher proficiency levels. The analysis of the variable data is couched within a stochastic version of the framework of Optimality Theory (Boersma's 1998 Gradual Learning Algorithm) because it allows the encoding of variability and its frequency effects within a language by means of a single grammar. Moreover, this study promotes a multidisciplinary and integrative approach that combines theoretical and methodological tools from three linguistic disciplines: sociolinguistics, second language acquisition and formal phonology, in an attempt to develop a "socially realistic linguistics" (Wilson & Henry, 1998).

Acknowledgments

This thesis would not have been possible without the help and guidance of a large number of people. First of all, I would like to thank my advisor, Dr. Walcir Cardoso, for his patience, continuous support, and for all the hours spent on the improvement of this thesis. I would also like to acknowledge the financial support as a research assistant to Dr. Cardoso from an FQRSC grant (NC-96880). It has been a true privilege to work for him and to have him as my supervisor.

Many more people have contributed to this thesis in one way or another. I am grateful for the support by the members of the thesis committee, Dr. Elizabeth Gatbonton and Dr. Pavel Trofimovich. They both have been very encouraging during this process. I would also like to express my gratitude to Professor Joanna White for giving me the opportunity to become a student in the Applied Linguistics programme at Concordia University.

I am grateful to the audience at the XIX Journées de Linguistique at Laval University for their insightful comments. They have contributed a great deal to the improvement of this thesis.

My special thanks go to Almudena Sainz, Director of language institute Interlingua and Raul Cervantes, Academic Coordinator at Interlingua. Their help and great attention to detail made the data collection possible. I am forever indebted to the students in Mexico who willingly participated in this study. Without their help, this investigation would not have been possible.

I would like to show my appreciation to Dr. Darin Howe for starting my interest in phonology. Without his extraordinary lectures at the University of Calgary I would not

be here at this point presenting research in the field. Thanks to Dr. John Archibald for his unconditional support and guidance in the transition from the Linguistics to the Applied Linguistics program. I miss the fruitful discussions about Second Language Acquisition and Phonology.

Finally, I am deeply grateful to my parents, Mario and Martha Escartín for their financial and moral support. Their encouraging words have motivated me to work hard and achieve my goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS/ITEMS

LIST OF FIGURES

LIST OF TABLEAUX

1. INTRODUCTION

The speech of a second language learner is a system composed of features of the learner's first language (L1), the second language (L2), as well as of principles that describe the core grammar of all languages (i.e., Universal Grammar – UG) (Major, 2001). Such system is commonly referred to as Interlanguage (Selinker, 1972). A learner's Interlanguage (IL henceforth) is variable in the sense that it is characterized by the production of both target and nontarget-like grammatical forms or sounds. For instance, the IL speech of Hispanophone learners of English is characterized by the variable application of [e]-epenthesis in words containing s+consonant (sC henceforth) clusters (e.g., 'steak' /stejk/ → [es.tejk], 'slim' /slɪm/ → [es.lɪm]).

Phonological variability such as the one described above has been usually accounted for via variable rules (e.g., Labov, 1969; Dickerson, 1975; Fasold, 1984), which serve to specify the linguistic environments in which phonetic variants are more likely to occur. To exemplify, assume that the production of sC clusters can be described via the variable rule illustrated below, where the variable application of [e]-epenthesis is indicated in parentheses. The subscripted letters *a* and *b* regulate the frequency of epenthesis, and the environment that has the most effect is associated with *a*.

(1) A variable rule for English sC clusters

$$\emptyset \rightarrow ([e]) \; / \; \left\{ \begin{array}{l} [+\text{consonantal}]_a \\ [-\text{consonantal}]_b \end{array} \right\} \; \_ \; [s]$$

(adapted from Carlisle, 1991a)

According to this rule, the environment [+consonantal]$_a$ (i.e., consonants preceding [s]) has the most effect on the application of [e]-epenthesis, while the environment [-consonantal]$_b$ (i.e., vowels preceding [s]) has a lower effect on the phenomenon. The problem with this rule is that, if one wishes to incorporate other phonological environments such as those following [s]: nasals, liquids and stops, then new rules need to be established and a higher level of complexity is added to the rules. Specifically, new rules need to be specified for each environmental factor, something which is not constrained. More recent studies have acknowledged the necessity of adopting new approaches for the analysis of variation such as the framework of Optimality theory (Reynolds, 1994; Anttila, 1997; Cardoso, 2001, 2004) because it allows one to account for variability via a simpler mechanism: a single constraint-based grammar.

Variability in IL is not only influenced by linguistic factors such as preceding or following phonological environments but also by social (Beebe, 1977; Schmidt, 1977), and psycholinguistic factors (Krashen, 1981; Tarone, 1985). Some recent studies have acknowledged the necessity of conducting more comprehensive investigations of IL variability by adopting a sociolinguistic methodology for data collection and analysis, and thus incorporating multiple variables motivated by factors from different disciplines (Young, 1991; Preston & Bayley, 1996; Cardoso, 2004). This allows one to obtain a more realistic view of the linguistic and extralinguistic factors that exert influence on IL variability. Interestingly, the number of studies that have incorporated both current developments in phonological theory as well as sociolinguistic methodology for data collection and analysis are still infrequent.

Cardoso (2004) investigated the variable speech of Brazilian Portuguese learners of English by following a sociolinguistic methodology for data collection and analysis, and by incorporating tools from three different disciplines: Second Language Acquisition, Phonology and Sociolinguistics. The study analyzed variation by adopting a stochastic version of the framework of Optimality Theory, and more precisely, the Gradual Learning Algorithm (OTsoft; Boersma, & Hayes, 2001). Through the adoption of a multidisciplinary and integrative approach to data collection and analysis, he has provided "a more comprehensive analysis of variation in the speech of second language learners" (Cardoso, 2004: 1). This study motivated the adoption of a similar approach to carry out a more thorough investigation of the variable speech of Hispanophone learners of English.

The goal of the present thesis is to investigate the variable development of /s/ plus consonant onset clusters in the speech of Hispanophone EFL learners across three levels of proficiency. These sequences pose difficulties for Hispanophone learners of English because they are disallowed in their L1, and because some sC clusters violate linguistic principles such as sonority sequencing. During the initial stages of acquisition, learners make use of [e]-epenthesis in order to syllabify all sC clusters (e.g., 'snake' /snejk/ → [e̲s.nejk], 'slim' /slɪm/ → [e̲s.lɪm]). Nonetheless, the application of [e]-epenthesis before s+liquid, s+nasal and s+stop onset clusters is variable. As we will see in Chapter 2, previous studies on the development of sC clusters have not provided a comprehensive examination of the phenomenon because: (1) they have not incorporated a large range of linguistic and extralinguistic factors, (2) their analyses have not been supported by phonological theory (e.g., resyllabification, markedness), and (3) they have not

incorporated sociolinguistic methodology for collecting and analyzing data that is intrinsically variable. In light of the necessity to carry out a more comprehensive investigation of the variable patterns that characterize [e]-epenthesis in the acquisition of L2 English, this study incorporates linguistic factors such as preceding phonological environment (i.e., vowel, consonant, pause, the former of which may trigger resyllabification across words), onset cluster markedness, and extralinguistic factors such as proficiency and level of formality.

The present study ventures to make an innovative contribution to L2 research by adopting a variationist perspective and a set of linguistic and extralinguistic factors to investigate the development of sC clusters in the speech of Hispanophone EFL learners. As we will see in subsequent chapters, an important contribution is made by investigating variability using current advances in phonological theory such as Optimality Theory, and by incorporating sociolinguistic methodology for data collection and analysis. This integrative approach to L2 research incorporates a more comprehensive set of factors to investigate the aforementioned phenomenon and identifies the possible interactions between the linguistic and extralinguistic factors that exert influence on L2 phonological acquisition.

The thesis is organized in the following way: Chapter 2 provides a description of the English and Spanish syllable structures because it is at this prosodic level that English sC clusters pose difficulties for Hispanophone EFL learners. This is followed by a discussion of how previous approaches such as syllabification via appendix have not been able to account for the syllabification of English sC clusters. This thesis argues that the framework of Optimality Theory is better able to justify the sonority violations of sC

clusters via a relative violability of constraints. Then a discussion of some previous studies that have incorporated linguistic and/or extralinguistic factors in the analysis of L2 phenomena is provided. This discussion leads to the suggestion that a multidisciplinary approach to data collection and analysis will provide a more comprehensive analysis of the development of sC clusters.

Chapter 3 describes the sociolinguistic methodology for data collection and analysis adopted in this thesis. More precisely, a description of the participants recruited for the investigation, and the instruments designed for data collection is provided. Subsequently, the steps involved in the data gathering procedure, the data recording and transcription are described.

Chapter 4 provides a brief introduction to Goldvarb 2001, the statistical program adopted for the multivariate analysis of the collected data. Following the introduction, the chapter presents and discusses the results from the linguistic and extralinguistic factors included in this investigation. This thesis shows that variation in the acquisition of sC onset clusters is conditioned by a combination of linguistic: sC sonority and preceding phonological environment, and extralinguistic factors: proficiency.

Chapter 5 provides a general introduction to the framework adopted for the analysis of the variable data: Optimality theory. Variation in Optimality Theory is described within three approaches: the multiple grammars approach, the crucial nonranking of constraints, and Stochastic OT – the approach adopted in this investigation. The chapter ends with the analysis of the variable production of s+liquid, s+nasal, and s+stop clusters within a stochastic version of the framework of Optimality Theory.

Finally, Chapter 6 addresses the implications of this thesis for Second Language Acquisition and Language Teaching, and the limitations that deserve consideration in future studies. This will be followed by my concluding remarks to the thesis.

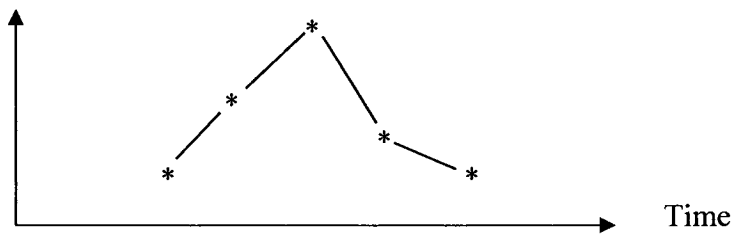
2. THEORETICAL FRAMEWORK AND RESEARCH QUESTIONS

2.1 Introduction to Syllable Structure

The application of [e]-epenthesis to English words (e.g., /slɪm/ → [es.lɪm]) involves the prosodic domain of the syllable, therefore an overview of syllable structure and its constituency is essential in order to arrive at a better understanding of how English sC onset clusters develop in the IL phonologies of Hispanophone EFL learners. Let us start by providing a definition of the syllable structure, which will be followed by a discussion of the syllable structures of Spanish and English.

A syllable consists of a prominent or sonorous peak (usually a vowel), surrounded by consonants that decrease in sonority towards the edges. Consider the example provided below, adapted from Giegerich (1992).

(2) Syllable structure and sonority



[k l æ m p]

Observe that sonority is related to syllable structure. The tendency to abide to sonority described above is called the Sonority Sequencing Generalization (Selkirk, 1984). Sonority sequencing is a universal principle of UG governing sequences of sounds, which seems to hold for all languages, in one form or another.

Note in (2) that the syllable structure of the English word 'clamp', exhibits a continuous rise in sonority towards the peak (i.e., the vowel [æ]) and decreases in sonority towards the edges. The types of consonants that can occupy the edge positions are determined by the Sonority scale (see Table 1 below). According to the sonority scale below, in the word [klæmp], [k] is less sonorous than [l], which is in turn, less sonorous than [æ]. Accordingly, in the consonants following the peak [æ], [m] is more sonorous than [p].

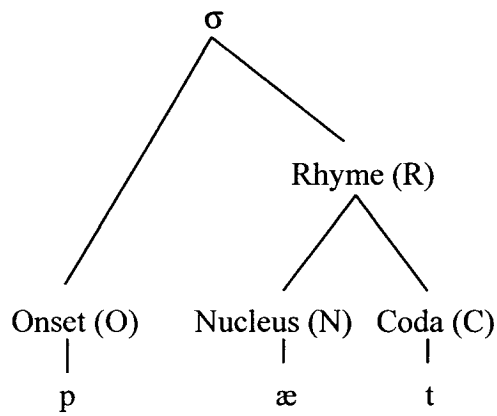
Table 1. Sonority Scale

Stops >	Fricatives and Affricates >	Nasals >	Liquids >	Glides >	Vowels
p	f	m	l	j	a
t	θ	n	r	w	ɔ
k	s	ŋ			i
b	ʃ				u
d	v				æ
g	ð				ɑ
	z				ɪ
	tʃ				e
	dʒ				o
less sonority/ more sonority					

Syllables have an internal structure composed of the following constituents: onset, and rhyme, the latter of which comprises the nucleus and coda. The nucleus is the only obligatory element in the syllable, and it is the most sonorous (i.e., a vowel), the consonant preceding the nucleus is called onset and it is optional. The consonant

following the nucleus, which is also optional, is called coda. The nucleus and the coda are contained within the rhyme, as exemplified in (3).

(3) The structure of the syllable



Languages differ in the number and kinds of segments that they allow in onset and coda positions. In order to arrive at a better understanding of the differences between English and Spanish syllable structure and how such differences may influence [e]-epenthesis, one needs to consider the relevant features that characterize this prosodic domain in the two languages. The following section reviews the syllable structure of Spanish, as described by Harris (1989).

2.2 Spanish syllable structure

Spanish syllables cannot contain more than five segments, which limits the number of segments allowed in onset and coda positions. Singleton onsets (i.e., one-member onsets) can contain any [+consonantal] segment in the phonemic inventory illustrated in Table 2 below.

Table 2. Spanish [+consonantal] phonemes

	Bilabial	Labiodental	Dental	Interdental	Alveolar	Palatal	Velar
Stop	p b		t d				k g
Affricate						tʃ	
Fricative		f			s z		x ɣ
Nasal	m				n	ɲ	
Liquid					l r		

In Spanish, it is then possible to observe words such as [**p**a.sa] 'raisin', [**t**a.ko] 'taco', [**s**a.la] 'living room', [**f**o.ko] 'light bulb'. It should be noted, however, that /ɲ/ only occurs word-internally (e.g., [ni.ɲo] 'boy', [ma.ɲa.na] 'tomorrow').

Spanish phonotactic constraints allow maximally two segments in onset position. Such sequences can only be formed by an obstruent [p, t, k, b, f, d, g] followed by a liquid [l, r]. Examples of possible complex onsets are provided in Table 3.

Table 3. Spanish complex onsets

Obstruent	Liquid	
	l	r
p	[plato] 'dish'	[prisa] 'hurry'
t	[tlatelolko] 'Tlatelolco'	[truko] 'trick'
k	[klasiko] 'classic'	[kreo] 'I believe'
b	[blusa] 'blouse'	[brazo] 'arm'
d	Ø	[dragon] 'dragon
g	[glukosa] 'glucose'	[grasa] 'fat'
f	[flor] 'flower'	[frasko] 'jar'

Observe that onset clusters of the type /dl/ are disallowed in Spanish, as illustrated in Table 3. In addition to this, s + consonant clusters are disallowed in this language (e.g., *[**sp**a.tu.la] ~ [**es.p**a.tu.la] 'spatula', *[**sl**a.βon] ~ [**es.l**a.βon] 'link of a chain'; see forthcoming discussion in section 2.4). Consequently, sC clusters can only occur when preceded by the vowel /e/ because it allows the syllabification of the first segment /s/ as the coda of the preceding syllable, and the second segment as the onset of the following one (e.g., [**es.p**a.tu.la] 'espatula').

The Spanish rhyme can contain from one to three segments. The first segment is a vowel, which can be followed or preceded by a glide (e.g., [m**uj**] 'very', [d**jo**] 'he/she gave'). The second and third segments are codas, the third of which can only be [s] (e.g., [o**bs**.ta.ku.lo] 'obstacle').

In sum, the syllable structure of Spanish strictly follows the sonority scale (Harris, 1983): the segments contained in the onset always rise in sonority towards the nucleus, while those in the coda decrease in sonority away from the nucleus. As we will see in the following section, English syllable structure is more complex than that of Spanish as it allows certain clusters that do not abide to the sonority profile, that is, s+stop [p, t, k] onset clusters.

2.3 English syllable structure

The syllable structure of North American English is more complex than that of Spanish because it can contain up to eight segments (Celce-Murcia, Brinton, & Goodwin, 1996; Spencer, 1996). Let us start the present section by describing the segments allowed in onset position. Onsets can consist of one to three segments. Almost all [+consonantal]

segments in the inventory (except /ŋ/ and /ʒ/) can syllabify as singleton onsets. Table 4 below illustrates the NAE consonant phonemes.

Table 4. English [+consonantal] phonemes

	Bilabial	Labiodental	Interdental	Alveolar	Palatal	Velar	Glottal
Stop	p b			t d		k g	
Affricate					tʃ dʒ		
Fricative		f v	θ ð	s z	ʃ ʒ		h
Nasal	m			n		ŋ	
Liquid				l r	r		
Glide	w				j		

Most two-segment onsets consist of sequences of a stop + a liquid (e.g., '**bl**ouse', '**gr**eed', '**cr**ab'). Additionally, English allows s+liquid, s+nasal and s+stop onset clusters, as illustrated in Table 5.

Table 5. sC English onsets

s + liquid		s + nasals		s + voiceless stops		
l		m	n	p	t	k
sly		**sm**all	**sn**ake	**sp**eak	**st**op	**sk**ate

Observe that the only segment that can precede a nasal or a voiceless stop in onset position is [s]. Equally important is the fact that English allows s+stop onset sequences even though these clusters violate the sonority profile illustrated in (2) and Table 1 above. This issue will be further addressed in section 2.4. Finally, three-segment onsets consist of a sequence of [s] + voiceless stop + the liquid [r] (e.g., '**spr**ay', '**scr**eam', '**str**ange'), except [sp] which can also be followed by [l] (e.g., **spl**ash).

English rhymes can include from one up to four segments. Two-segment rhymes consist of a vowel followed by any [+consonantal] segment. Three-segment rhymes contain a vowel followed by a sequence of either a nasal+obstruent (e.g., 'lump' [lʌ**mp**]) or a liquid+obstruent (e.g., 'gulp' [gʌ**lp**]). Finally, four-segment rhymes are only found in morphologically complex structures (e.g., 'texts' [tɛ**ksts**]).

From the descriptions of Spanish and English syllable structures, observe that English and Spanish allow both complex onsets and complex codas. While English onsets can contain up to three segments, codas can comprise up to four segments. Spanish, on the other hand, allows maximally bilateral (two-member) onsets and bilateral codas. Furthermore, English allows s+consonant (i.e., s+stop, s+liquid, and s+nasal) onset clusters, something which Spanish phonotactics disallows. The following section will focus on the syllabification of sC onset clusters in Spanish based Interlanguages.

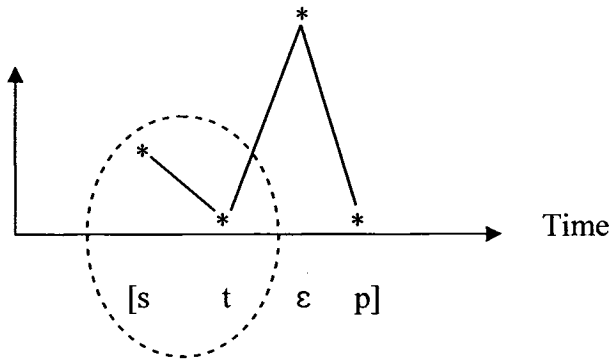
2.4 sC syllabification in Spanish based Interlanguage

As indicated in previous sections, Hispanophone learners of English have difficulties in the production of s+nasal and s+liquid clusters (all of which abide to the sonority profile; see (2) and Table 1 above) and s+stop clusters (which violate sonority

sequencing – see forthcoming discussion under (4) below). This suggests that these learners merely transfer the constraints on sC clusters from their L1 by breaking the illicit cluster with an epenthetic [e]; e.g., 'snake' /**sn**ejk/ → [es.**n**ejk], 'slim' /**sl**ɪm/ → [es.**l**ɪm], 'stop' /**st**ɑp/ → [es.**t**ɑp]. The production of sC clusters in interlanguage, however, is not categorical as the discussion above implies. Variability among s+stop, s+nasal, and s+liquid onset clusters may be triggered by a combination of both extralinguistic (e.g., level of formality, proficiency) and linguistic factors (e.g., preceding phonological environment, sonority profile of the two members of the cluster). For instance, under the concept of markedness, s+stop onset clusters are considered marked (i.e., uncommon or "more difficult to produce") because they violate the sonority profile discussed in section 2.1, while s+nasal and s+liquid clusters are considered unmarked (i.e., common or "easier to produce") since they abide to sonority (Eckman, 1977). Accordingly, if a language allows the more marked structure (i.e., s+stop onset clusters), then it will also allow the least marked structure (i.e., s+nasal, s+liquid onset clusters). Because marked structures are more difficult to acquire, I hypothesize that Spanish learners of English will acquire s+liquid and s+nasal clusters before s+stop clusters (see also previous studies in section 2.5.1).

If sonority sequencing within the syllable is a universal principle, how can one explain its violation in languages that allow s+stop onset clusters (e.g., English, French)? As discussed in previous sections, sonority within the syllable must increase towards the nucleus and decrease away from it. English s+stop onset clusters, however, violate sonority sequencing within the syllable because they contain segments that both decrease and increase in sonority towards the nucleus, as exemplified in (4).

(4) A violation of Sonority



In previous theoretical frameworks in which constraints are deemed inviolable, researchers have assigned [s] to an extra-syllabic margin or appendix (e.g., Giegerich, 1992; Harris, 1994). An appendix is directly integrated into the phonological hierarchy at the word level (i.e., the Prosodic Word – PWd) without incurring in violations of sonority, and avoiding intermediate levels of constituent structure such as the onset.

(5) /s/ syllabification via appendix

It should be noted, however, that only /s/ can prosodize as a syllable appendix, and that appendices can only occur before /p, t, k/ (Giegerich, 1992). The concept of appendix is an ad hoc solution because there does not seem to be either a theoretical or an empirical motivation to attach /s/ to a higher constituent other than to provide a justification for the violation of sonority in s+stop onset sequences. In the framework of Optimality Theory (OT) adopted in this investigation, the notion of appendix is unnecessary because constraints may be violated in order to satisfy higher ranked constraints (e.g., a constraint requiring the input segment /s/ to surface in the output). This theoretical framework will be discussed in Chapter 5.

For a comprehensive explanation of [e]-epenthesis in the variety of Spanish/English under study, one needs to consider not only markedness but also other linguistic factors (i.e., preceding phonological environment and its effect on resyllabification) and extralinguistic factors (i.e., level of formality and proficiency level), as they may have an effect on the development of sC clusters. The following section will discuss some previous studies, which have incorporated linguistic and/or extralinguistic factors in the analysis of L2 phenomena.

2.5. Previous studies

2.5.1 Linguistic effects on Interlanguage phonology

The present section addresses some linguistic factors that may influence the development of patterns of variation in Interlanguage. In the context of Spanish, Carlisle (1991a) carried out a study with 5 Hispanophone ESL learners. The objective of the study was to determine if there is a greater frequency of epenthesis in bilateral onsets that

violate the sonority profile described in section 2.1. Specifically, Carlisle (1991a) expected Hispanophone ESL learners to apply [e]-epenthesis more frequently before s+stop onset clusters because according to the markedness hypothesis (Eckman, 1977) clusters that violate the sonority profile are considered rare phenomena (i.e., marked), and are therefore expected to be more difficult to acquire. It was found that epenthesis occurred more frequently before /st/ onset clusters and less frequently before /sl/ and /sm/ onset clusters. This finding confirms the prediction that the less marked onsets clusters surface earlier and more frequently than the more marked ones. In addition to sonority and markedness as influencing factors, Carlisle's (1991a) study showed that the proportion of epenthesis was greater after consonants than after vowels. However, he provided no further explanation as to why consonantal environments induced a higher frequency of epenthesis. A tentative explanation for this is that an original vowel preceding an sC onset cluster (as opposed to a preceding consonant) can be resyllabified as the nucleus of a new syllable thereby leading to a lower frequency of epenthesis after vocalic environments (e.g., /ə stæmp/ → [əs . tæmp]). Carlisle also suggested that the sonority of the preceding consonantal environment may influence the frequency of epenthesis. His study in fact showed that preceding obstruents induced the highest frequency of epenthesis, while vowels (the most sonorous segments) induced the lowest. More precisely, Carlisle (1991a) suggested that "the degree of sonority of the preceding environment may be the true determinant of the frequency of epenthesis" (p. 90). In order to investigate such a claim, the present study included an equal number of preceding consonantal environments (i.e., an equal number of preceding obstruents and consonantal sonorants) in the formal task. This will be further discussed in Chapter 3.

In a similar line of research, Carlisle (1991b) investigated the occurrence of epenthesis involving only s+stop clusters (i.e., /sp/, /st/, and /sk/). Results showed a higher frequency of epenthesis after consonantal environments but no significant differences in the frequency of epenthesis between the three onsets.

In these studies, Carlisle made use of variable rules to account for the linguistic factors (i.e., preceding phonological environment and sC onset sonority) that significantly contributed to the application of [e]-epenthesis. Nevertheless, his studies did not incorporate recent advances in phonological theory such as a stochastic version of OT to account for both the application of [e]-epenthesis and the correct production of sC clusters (see introduction to stochastic OT in Chapter 5).

In addition, his studies did not take into account either the development of sC clusters overtime or extralinguistic factors such as levels of formality and proficiency, all of which have been shown to have an effect on variability in Interlanguage (Major, 2001; Major, 2004; Cardoso, 2004; John, to appear. See also forthcoming discussions). It is therefore necessary to consider the interrelation between preceding environment (which may trigger resyllabification), sonority, proficiency and formality for a more comprehensive investigation of the development of sC onset clusters in the speech of Hispanophone EFL learners.

One of the first studies that acknowledged the necessity of adopting a more current theoretical framework to analyze L2 data (i.e., Optimality Theory) was the one carried out by Hancin-Bhatt and Bhatt (1997). Their study examined the interaction between cross-language transfer effects (i.e., L1 to L2 transfer) and developmental effects in the production of complex syllable onsets in a group of Japanese and Spanish ESL

learners. Their study predicted that cluster acquisition is constrained by L1 sequencing possibilities and by universal constraints on sequencing. However, results showed variation in error types, something that contradicted the researchers' predictions. The authors concluded that Optimality Theory is an ideal framework to capture the factors intervening in Interlanguage grammars such as L1 to L2 transfer, markedness, and sonority. Specifically, OT indicates which structures are illicit in a particular language, and accounts for how and why illicit structures surface the way they do, conforming to the grammar available to the speaker.

Cutillas-Espinoza (2002) made the first attempt to account for [e]-epenthesis in Spanish English by means of Optimality Theory. The framework was used to explain that, even though both s+stop and s+liquid onset clusters are disallowed by Spanish syllable structure, the former appears to be more difficult. The higher percentage of s+liquid sequences indicated that they were produced more successfully and, accordingly, the lower percentage of s+stop sequences suggested that these clusters are more difficult for upper-intermediate EFL learners. Although such results are consistent with the findings provided by previous studies (Carlisle, 1991a; 1991b; Hancin-Bhatt & Bhatt, 1997), they do not seem to provide strong evidence to support the author's claims that s+stop clusters are more difficult to produce than s+liquid clusters. More precisely, the five participants included in this study showed only 4 instances of [e]-epenthesis in the data (i.e., 3 instances of [e]-epenthesis before s+stop clusters and 1 before s+liquid clusters). This suggests that the production of sC clusters could have been influenced by the high proficiency level of the participants involved in this study. Specifically, time and increased exposure to the L2 could have significantly contributed to the correct

production of sC clusters. It seems that a cross-sectional study might provide a clearer pattern of decreasing epenthesis.

In addition, the study provides us with no information as to the variable patterns involved in the acquisition of sC onset clusters and no information as to how formality levels or preceding phonological environments influence the frequency of [e]-epenthesis. It then appears relevant to take into account variability, development over time and formality in future investigations of [e]-epenthesis in order to obtain a more accurate view of the factors (i.e., linguistic and extralinguistic) that influence the production of sC onset clusters.

The first study that investigated if the patterns of [e]-epenthesis changed over time as L2 proficiency increased was the one carried out by Abrahamsson (1999). This study replicated the various studies reported by Carlisle in order to find out if the patterns demonstrated with elicited speech also held for natural speech data in a longitudinal case study of one L1 Spanish learner of Swedish[1]. The data collection included nine half-hour studio recordings made at approximately 3 to 5 week intervals from August 1990 to May 1991 plus one follow-up recording in March 1992. The data recordings consisted of conversations about everyday topics, newspaper articles, world events, and political events between the participant and two native Swedish interviewers. Results showed that bilateral onsets were epenthesized less frequently than trilateral ones, and that there was more epenthesis after consonantal environments. The sonority relationship among onset members revealed that, contrary to what was predicted[2], /sl/ onset clusters were

[1] The study is relevant for the present discussion because Swedish syllable structure allows initial consonant clusters that violate the sonority profile (just like English): e.g., [sp, st, sk, spr, skr, str, spj, spl].

[2] It was predicted that epenthesis would be higher in sC clusters that violate sonority (e.g., [sp, st, sk]).

epenthesized more frequently than s+stop and s+nasal clusters. Abrahamsson offers, however, no explanation as to why these clusters behaved in such a way. He merely mentions that this speaker does not conform to the pattern. In addition, the learner's development over time did not show the expected decline from high to low frequencies of epenthesis. Instead, it showed a pattern of low-high-low frequencies of epenthesis. Abrahamsson's (1999) explanation for this unpredictable pattern was that increased L2 proficiency allowed the learner to produce less attended speech. It should be pointed out that this unexpected pattern of development contradicts the Ontogeny Phylogeny Model (Major, 2001), which predicts that L1 transfer features (e.g., [e]-epenthesis) decrease overtime, while L2 features increase (e.g., sC clusters). In addition, Abrahamsson's (1999) study provides us with no information about the participant's knowledge of other languages (e.g., English). Specifically, if the participant was learning Swedish as a third language, she might have transferred her knowledge of English bilateral and trilateral onset clusters to Swedish. In addition to this, the information obtained from longitudinal case studies involving one single participant may not always be representative of a larger group.

The results provided by the studies discussed in this section suggest that it is necessary not only to account for the variable production of sC onset clusters, but also to conduct a more thorough investigation of the phenomenon to determine which linguistic and extralinguistic factors influence the development of sC clusters in Interlanguage.

This goal can be achieved partly by: (1) adopting a more current framework in phonological theory (i.e., Stochastic Optimality Theory)[3], and (2) through the incorporation of a larger set of participants within different levels of proficiency because they might provide us with a clearer pattern of decreasing L1 transfer. Let us now consider the extralinguistic factors that can aid us in achieving the abovementioned goal.

2.5.2 Extralinguistic effects on Interlanguage phonology

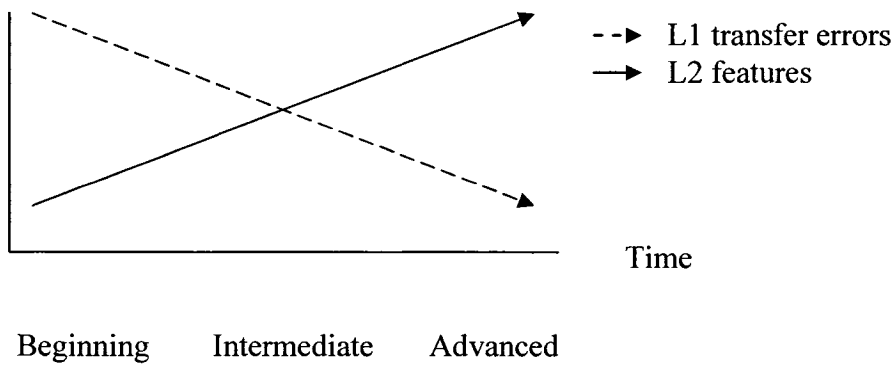
While the abovementioned studies focused on the effects of linguistic factors (i.e., markedness on sonority, and preceding environment) on L2 phonology, other studies have shown that the acquisition of a second language may also be influenced by extralinguistic factors such as formality levels, gender, interlocutor, proficiency level, etc. For instance, a number of sociolinguistic studies dealing with nonnative speaker variation have found that the frequency of standard forms increases in more formal situations mainly because the L2 speaker monitors speech more closely, thereby reducing the tendency for L1 transfer (Dickerson, & Dickerson, 1977; Gatbonton, 1978; Major, 2001; Cardoso, 2004; Major, 2004; John, to appear). In the context of Hispanophone learners of English, one would expect to find fewer occurrences of [e]-epenthesis in formal tasks and a higher frequency of the phenomenon in informal tasks.

Another extralinguistic factor that may influence L2 acquisition is proficiency. The proficiency level of an L2 learner may in turn be shaped by the increasing exposure

[3] A number of studies (Hancin-Bhatt & Bhatt, 1997; Cutillas-Espinoza, 2002) have adopted a standard version of Optimality Theory to account for the illicit structures in the speech of Hispanophone learners of English. Nonetheless, standard OT cannot account for the variable production of sC onset clusters without resorting to multiple grammars. As we will see in Chapter 5, Stochastic OT allows the encoding of variation and frequency of output forms within one single grammar.

to the second language. Recall from section 2.5.1 that Abrahamsson's (1999) study investigated the development of sC onset clusters over time. His study predicted that the frequency of [e]-epenthesis would decrease over time as L2 proficiency increased. Such a prediction is captured by the Ontogeny Phylogeny Model (OPM henceforth). The OPM foresees that at the beginning stages of L2 acquisition, the L1 exerts a strong influence. Such influence decreases with increased exposure to the L2. More precisely, L1 transfer errors (e.g., [e]-epenthesis) occur at a considerably high rate at the beginning stages and decrease as learners become more proficient in the L2 (i.e., overtime). The L1 transfer patterns and L2 features in the development of a second language are illustrated in (6).

(6) The OPM: L1 transfer patterns and L2 features



According to the OPM, L1 transfer errors decrease overtime but developmental errors (i.e., phenomena that are neither part of the L1 nor L2) increase and then decrease. Because the Hispanophone EFL learners in the studies discussed thus far do not exhibit developmental errors in the acquisition of sC clusters (i.e., their speech is characterized by the presence or absence of an epenthetic /e/), this study assumes that [e]-epenthesis is attributable to transfer despite also being triggered by general linguistic principles such as

sonority. Based on the studies discussed in this section, this study predicts that beginning learners will exhibit a high frequency of [e]-epenthesis, which will decrease with increased exposure to the L2 (Carlisle, 1991a; 1991b; Abrahamsson, 1999; Cutillas-Espinoza, 2002). The production of sC clusters as such (i.e., L2 features), on the other hand, is expected to be more prominent in more advanced groups of EFL learners.

It is clear that in order to carry out a more comprehensive investigation of the linguistic and extralinguistic factors that influence the variable application of [e]-epenthesis, a multidisciplinary and integrative approach to data collection and analysis must be adopted. Cardoso (2004) adopted such an approach in a study that incorporated theoretical and methodological tools from sociolinguistics, SLA and formal phonology. The study incorporated a set of linguistic factors (i.e., word status, place of articulation of the word-final stop, word size, and stress placement within the word) and extralinguistic ones (i.e., level of proficiency, style, and subjects) to investigate the variable acquisition of English singleton word-final codas by Brazilian Portuguese learners of English. This innovative study analyzed variation within a stochastic version of OT (i.e., via the Gradual Learning Algorithm proposed by Boersma et al, 2001) because it reveals whether a grammar is categorical or variable, and more importantly, it allows the encoding of frequency of output forms into the grammar. In the study, results showed that the occurrence of English singleton codas is higher in more formal stylistic environments and in higher levels of proficiency. Through the use of the Gradual Learning Algorithm (GLA), Cardoso (2004) made an important contribution by showing how "the GLA allows the encoding of variability and its frequency effects within a language by means of a single crucially ranked grammar" (p.10). This study motivated the adoption of a

similar approach to carry out a more thorough investigation of the variable speech of Hispanophone learners of English.

2.6 Research questions and hypotheses

As implied in the discussions above, previous studies on the development of sC onset clusters have not investigated the phenomenon from a more comprehensive variationist perspective. Given the fact that the production of these clusters in interlanguage is intrinsically variable, the present study will attempt to determine which factors (i.e., linguistic and extralinguistic) influence the extent to which Hispanophone EFL learners apply [e]-epenthesis.

Based on the discussions and the literature review provided in this chapter, the study will address the following research questions:

---

1. How do English sC onset clusters develop during the acquisition of EFL across three proficiency levels?
2. Does the sonority profile of the sC onset clusters have an effect on variability in [e]-epenthesis? Specifically, are the less marked onset clusters (i.e., s+liquid, s+nasal) more likely to surface earlier and more frequently than the more marked sequences (i.e., s+stop)?
3. What is the effect of preceding phonological environment (i.e., pause, vowel, and consonant) towards [e]-epenthesis?
4. How does level of formality influence the frequency of epenthesis in sC clusters?

The hypotheses that arise from the research questions are:

1. The frequency of [e]-epenthesis will decrease as L2 proficiency increases.
2. The less marked clusters (s+liquid, s+nasal) will surface earlier and more frequently than the more marked clusters (s+stop).
3. Preceding consonantal environments will induce the highest frequency of [e]-epenthesis.
4. There will be a higher frequency of [e]-epenthesis in less formal stylistic environments (i.e., in less formal tasks)

The following chapter describes the methodology that was followed in order to investigate the abovementioned research questions and hypotheses.

3. METHODOLOGY

This chapter provides a description of the methodology followed in this study: recruiting of participants, the rationale and description of instruments, data gathering, data recording and transcription. The chapter ends by addressing the process followed for the analysis of the collected data. Let us start by providing a description of the participants recruited for this study.

3.1 Participants

The participants were 23 Hispanophone learners of English across three proficiency levels: beginning, intermediate and advanced, whose ages ranged from 21 to 30. These participants were selected from EFL classes at a language institute in Mexico. Their proficiency level was initially determined through a brief oral interview, which included questions similar to those asked during the entrance interview at this language institute (see (7) below). Through this interview, it was verified if a student could use different grammatical structures (e.g., simple present, passive voice, present perfect) in affirmative, negative, and interrogative forms. The interviews comprised three tasks: answering questions, asking questions, and translating sentences from Spanish to English.

(7) Diagnostic interview: sample questions

> Does your father live in Coatzacoalcos?
> Can you tell me some information about the people in your family?
> Can you ask me some questions about the people in my family?
> Please tell me some places that exist near your home.
> Pregúntame de donde viene el tequila – 'ask me where tequila comes from'
> Pregúntame si hay libros en mi cuarto – 'ask me if there are any books in my bedroom'

Participants' level of proficiency was later confirmed by a verification of the oral grades obtained in the language course prior to the data collection[4]. The total number of hours of instruction also determined the participants' proficiency level: beginning learners – approximately 82 hours, intermediate learners – 137 hours, and advanced learners – approximately 275 hours of EFL instruction. Upon selection of participants and verification of their proficiency level, three experimental groups were established, as illustrated in (8).

(8) Distribution of participants in three experimental groups

Group 1	**Group 2**	**Group 3**
8 beginning learners of English	8 intermediate learners of English	7 advanced learners of English

3.2 Instruments

The present study included a total of three instruments. Specifically, participants filled out a background questionnaire and then performed both a formal and an informal task. The following sections provide descriptions and the rationale of the instruments included in this study

[4] Participants' oral grades cannot be displayed because of a policy on the part of the language institute. It should be noted that, from the 23 participants included in this study, 2 had a minimum passing oral grade of 6, 3 obtained the highest grade – 10, and the rest had an average grade of 8. These scores then suggest that participants overall pronunciation and intelligibility were within the average expected by the school.

3.2.1 The background questionnaire

This questionnaire included a total of 8 questions, which sought participants' general information such as age, gender, self-rated proficiency, degree of motivation to learn an L2, exposure to English outside the classroom, and time devoted to learning pronunciation. Each question provided participants with a set of answers from which they were asked to select the option that best described their situation (see Appendix A for a sample of the questionnaire used). Participants' motivation to learn English was included because previous studies have shown that motivation appears to be the second strongest predictor of language learning success after aptitude (Skehan, 1989). The information included in this questionnaire is relevant to the present study because it helps control for these variables, and can help the researcher determine if these factors may exert an influence on the pronunciation of sC onset clusters

3.2.2 The formal task

The formal task consisted of an adaptation of the grammaticality judgment task used by Lin (2003). Her grammaticality judgment task was designed to investigate variability in IL consonant cluster simplification strategies within a group of Chinese EFL speakers. The task included 20 pairs of sentences, each of which contained at least one target item for her study, and violated English grammatical rules such as verb agreement. The task included a total of 54 target items.

(9) Lin's (2003) Grammaticality judgment task

(a) His favorite color is blue, but he is a fleep person.

(b) His favorite color is blue, but he is a fleep people.

This study incorporated a grammaticality judgment task because as shown in Lin's (2003) study "by forcing the subjects to focus their attention on grammar rather than pronunciation, we can obtain more natural speech" (p.447). Another advantage of using this kind of task is that it allows the researcher to elicit the pronunciation of a target word without letting participants know the real purpose of the task. A sample pair of sentences is provided in (10) below. The target items are italicized here; however, they appeared in plain font in the version distributed to the participants.

(10) Grammaticality judgment task

Read aloud the sentence that you consider as grammatically correct.

1. (a) A quick *study* session before the test sounds good.

(b) A quick *study* session before the test sounds well.

The grammaticality judgment task (see Appendix B) included the target onset clusters /sl, sm, sn, sp, st, sk/ as well as an equal number of preceding phonological environments (i.e., pauses, vowels, and consonants) in order to reinforce construct validity (Seliger & Shohamy,1989)[5]. More precisely, this task included 24 preceding consonants (i.e., 6 nasals, 6 stops, 6 liquids, and 6 fricatives) in order to further explore

[5] It should be noted that construct validity cannot be controlled for during the informal interview because the researcher has no control on the number and types of sC clusters produced in a task that elicits spontaneous speech.

Carlisle's (1991a) observation that the sonority of the preceding consonantal environments influences [e]-epenthesis (see section 2.5.1). In addition to this, 18 preceding pauses (i.e., 3 pauses before each target cluster) were included to confirm Abrahamsson's (1999) claim that these environments have a neutral effect on the frequency of epenthesis. Finally, the task included 18 preceding vowels of different heights equally distributed (i.e., 6 high, 6 mid, and 6 low) because previous studies have shown that preceding vowels have a lowering effect on the frequency of [e]-epenthesis (Carlisle, 1991a; Abrahamsson, 1999). Each of the target clusters occurred 10 times: 4 times before consonants, 3 times before vowels and 3 times before pauses (see Appendix C). This task was performed in approximately 15 minutes.

3.2.3 The informal task

The informal task consisted of an interview in which participants were asked some questions in a relaxed way as if giving closure to the data collection session (see Appendix D for a larger sample of the questions used). The purpose of this task was to elicit spontaneous speech and minimize observer's interference, as proposed by Labov's (1972) observer's paradox: "the aim of linguistic research in the community must be to find out how people talk when they are not systematically being observed; yet we can only obtain these data by systematic observation" (p. 209). An attempt was made to overcome this paradox by involving participants in questions that related to different personal topics (see (11) below). This task lasted approximately 20 minutes.

(11) Informal interview: sample questions

What kind of work do you do?
What do you like to do in your free time?
What was the last movie you saw? What was the movie about?
Is there another country you would like to visit or live in?
Do you have any special interests or hobbies?

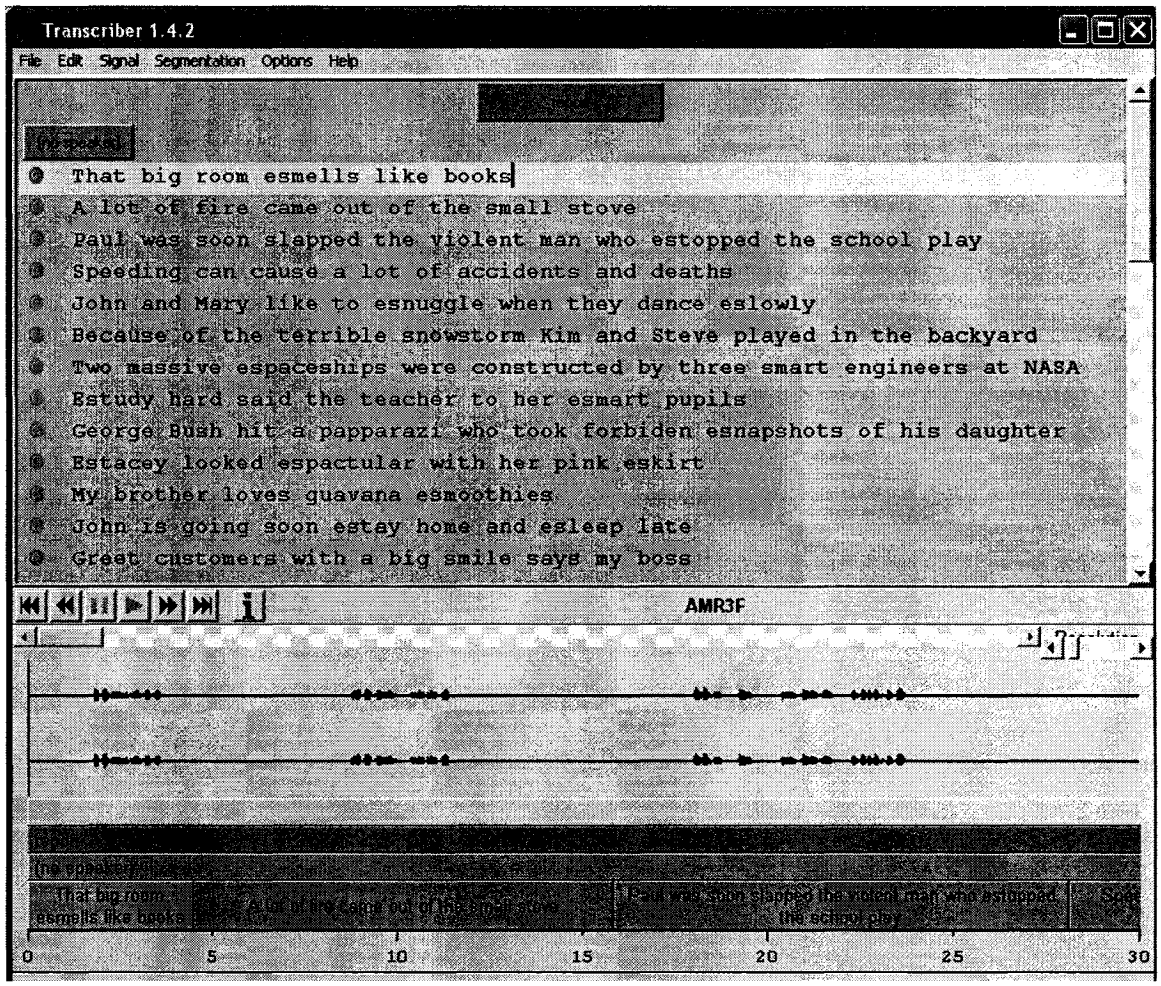
3.3 Data gathering procedure

The data gathering procedure took place at a language institute in Mexico. Participants booked an appointment with the researcher and were told that the tasks would be performed on an individual basis. During each session, the researcher explained the study without revealing its real purpose; participants signed a consent form and then proceeded to fill out the background questionnaire. The background questionnaire included a total of 8 questions to obtain the participants' general information described in section 3.2.1. During this task, participants answered the questions by marking the option that best described them and their learning situation. Following the questionnaire, participants were given 36 pairs of sentences printed on paper. They were told to select the option they considered as grammatically correct and then read it aloud. Participants had no time constraints during the completion of the grammaticality judgment task. Upon completion of the formal task, the researcher and each participant engaged in an informal conversation (i.e., the informal task). Each data collection session lasted an average of forty minutes.

3.4 Data recording and transcription

The two tasks were recorded via a Marantz CDR300 CD-R/RW recorder and an Audio-Technica AT831b lavaliere microphone for further transcription and coding. The collected data were later transcribed using Transcriber V1.21, a tool for segmenting, labeling and transcribing speech. A sample of a transcription file is provided in Figure 1 below. Observe in Figure 1 that, the upper part of the window shows the transcription of the sentences containing sC clusters. The presence of the epenthetic [e] was indicated with an orthographic "e" (e.g., "esmells" in line 1), while the sC words produced correctly were transcribed as such (e.g., "small" in line 2). The middle part of the window illustrates the wave forms of the transcribed sentences, which facilitates data manipulation. The program played the transcribed sentences while also allowing one to see the wave form of the sentences in question. The lower part of the window shows the corresponding segmented transcription.

Figure 1. Transcriber V1.21: Sample transcription file



Notice that the segmented transcription corresponds to the wave form of a specific sentence. This allows not only easy access to the transcribed sentences but also a rapid location of the wave forms for the sC clusters under scrutiny. When difficulties in identifying the presence or absence of an epenthetic /e/ arose, the data were checked with the program Praat (version 4.2.34, Boersma & Weenink, 2004), a program with which one can analyse, synthesize, and manipulate speech. Praat read the corresponding transcribed files, and provided audio playback along with the waveforms for the words in

question. A sample of the waveforms for the words [esmɛlz] 'esmells' and [smɛlz] 'smells' is presented in Figures 2 and 3.

Figure 2. Wave form for [esmɛlz] 'esmells'

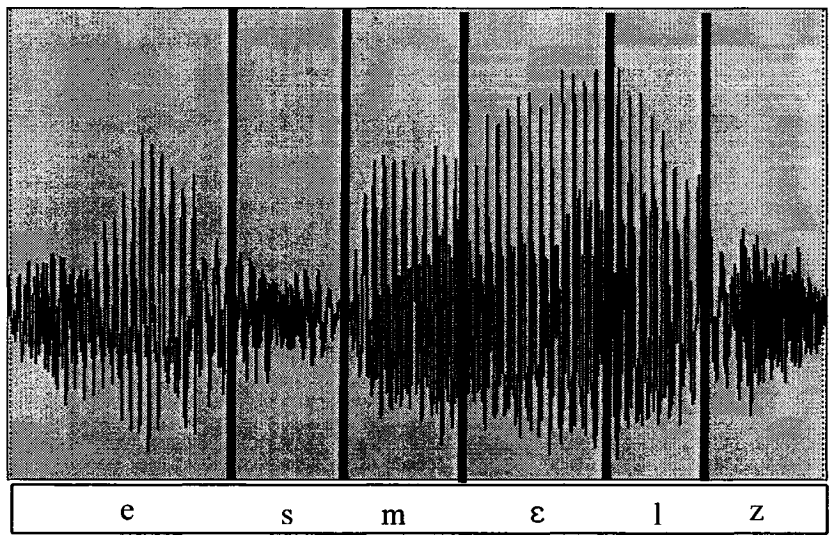
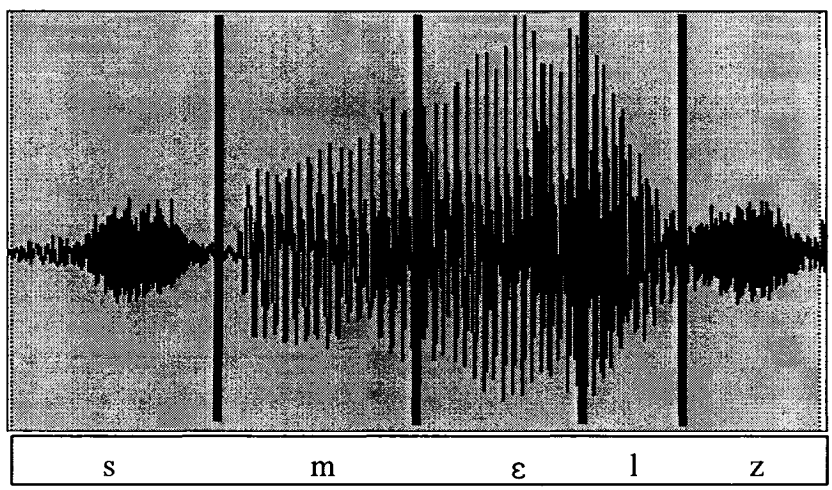


Figure 3. Waveform for [smɛlz]'smells'



Notice in Figure 2 that the pronunciation of 'smells' contains an epenthetic [e]. The high-amplitude periodicity at the beginning of this word indicates the presence of a segment with high sonority (i.e., a vowel). On the other hand, the wave forms in Figure 3 indicate that the pronunciation of the word 'smells' starts with a segment of low sonority (i.e., a fricative). It should be mentioned, however, that when the presence of an epenthetic [e] could not be identified with the abovementioned programs (i.e., Transcriber V1.41 or Praat), the tokens were rechecked in consultation with my advisor Dr. Cardoso.

3.5 Data analysis

The statistical program Goldvarb (Robinson, Lawrence, & Tagliamonte, 2001) was adopted in the present investigation because it allows one to analyze the kind of data that are collected in studies of Interlanguage variation involving spontaneous speech (Young, and Bayley, 1996). In contrast, a statistical analysis of variance (i.e., ANOVA) allows one to analyze data that come from controlled experiments. A brief introduction to Goldvarb will be provided in Chapter 4.

Upon completion of data transcription, the collected 1,527 tokens were stratified among five independent variables. Subsequently, a series of Goldvarb quantitative analyses were run to refine the model of variation (these kinds of analyses will be further explained in Chapter 4). Table 6 below shows the factor groups that were initially included in the analyses based on the hypotheses discussed in Chapter 2.

Table 6. Linguistic and extralinguistic factor groups for Goldvarb analyses

Factor group	**Factors**				
Dependent variables	Epenthesis	Target form			
sC sonority	s+liquid	s+nasal	s+stop		
Preceding environment	Consonant	Vowel	Pause		
Proficiency	Beginning	Intermediate	Advanced		
Level of formality	Formal	Informal			
Subjects	#1	#2	#3	#4	#5

In an attempt to further account for the variable production of sC clusters found in the speech of Hispanophone EFL learners, a series of analyses were carried, using a Stochastic version of the framework of Optimality Theory. The pertinent constraints were fed into a Gradual Learning Algorithm (i.e., OTsoft; Boersma, & Hayes, 2001), which assigns ranking values to the constraints. These ranking values determine whether a grammar is categorical or variable, and more importantly, serve to encode frequency effects into the grammar, unlike standard OT. The framework of Optimality theory will be further discussed in Chapter 5.

The following chapter provides a brief introduction to Goldvarb, followed by the presentation and discussion of the quantitative results obtained in this study.

4. GOLDVARB RESULTS

The main objective of the present study was to determine which factors influence the extent to which Hispanophone EFL learners apply [e]-epenthesis. In order to accomplish this goal, the collected 1,527 tokens were submitted to three Goldvarb 2001 analyses to obtain a more accurate representation of the phenomenon under investigation[6] (see Appendix E for the number of tokens per participant and Appendix F for the number of tokens across proficiency levels). The present section provides an introduction to this statistical program, which is followed by the presentation and discussion of the final results.

4.1 Introduction to Goldvarb

As mentioned in section 3.5, this study adopted the statistical program Goldvarb (Robinson, Lawrence, & Tagliamonte, 2001) not only because it is the most commonly used in sociolinguistics, but also because it handles the kind of data that are collected in studies involving spontaneous speech, that is, data or corpus that cannot be controlled for by the researcher (Young & Bayley, 1996). Specifically, this program allows one to conduct multivariate analyses to facilitate the understanding of the complex range of factors that may influence the choice of one variant over another and the systematicity that triggers variable language production.

The results obtained from a Goldvarb analysis can be interpreted as holding over the whole corpus under investigation and to all similar speakers and contexts. In order to arrive at a better understanding of the results obtained in this study, let us first discuss the

[6] A preliminary version of this study was presented at the XIX Journées de Linguistique (Université Laval) and will appear in Escartin (to appear). I would like to thank the audience for their questions and valuable comments, which have significantly contributed to the improvement of this thesis.

steps involved in a multivariate analysis in Goldvarb.

For the initial Goldvarb run, all words containing sC clusters were coded according to the scheme illustrated in Appendix G. According to this coding scheme, for instance, the sC initial word 'esmells' in 'that big room esmells like books' was coded as *1NnaF1*. The first character in the coding string (i.e., *1*) indicates the presence of an epenthetic [e], *N* represents the phonological environment following /s/ (i.e., a nasal), and *n* indicates that the preceding phonological environment is also a nasal. The last three characters in the coding string refer to the extralinguistic factors included in this study. Specifically, *a* denotes that the participant is an advanced English learner, *F* indicates that the token comes from the formal task, and *1* refers to the participant.

Upon completion of data coding, the next step was to create a cell file, in which the entered tokens are checked automatically to make sure that all characters represent legal values. The researcher then has to select the application value. For all runs of the tokens in the Spanish data, 1 (i.e., epenthesis) was selected as the application value. After creating the cell file, Goldvarb calculated the numbers and percentages of applications and non-applications as illustrated in Table 7 (see also Appendix H).

Table 7. Goldvarb results for the factor group *sC sonority*

Number of cells: 194
Application value(s): 1
Total no. of factors: 33

Group		Apps	Non-apps	Total	%
1 (2)					
N	N	153	310	463	30
	%	33	66		
S	N	331	482	813	53
	%	40	59		
L	N	101	150	251	16
	%	40	59		
Total	N	585	942	1527	
	%	38	61		

It should be noted that it is not possible to draw any conclusions from the raw number or percentage because they do not express the influence of each factor independently of the others. Before the actual multivariate analyses, it is advisable to use the cross tabulations command to check for interactions between factor groups (e.g., the percentage of [e]-epenthesis by proficiency and formality levels), and to identify the presence of categorical results. The identification of interactions between factor groups and categorical results will in turn allow the researcher to refine the model of variation by recoding tokens or regrouping factor groups. The next steps are to conduct the binomial 1-level, and binomial up and down statistical analyses. The binomial 1-level analysis shows the input probability of each independent factor (e.g., the likelihood of [e]-epenthesis to apply regardless of the presence or absence of any particular factor). This operation also shows the factor weight (*p*), which measures the influence that each factor has in the process under investigation. Additionally, it provides the most accurate insight into the likelihood of variant occurrence. The factor weight is a value associated with each factor independently of other factors in the same factor group. The higher the value (e.g., 1.00), the higher the influence of that factor in the selection of the variable output. More precisely, a value of 0.00 indicates that the variant will never appear, while a value of 1.00 indicates that the variant will always appear. Because the production of sC onset clusters involves two variants, the factor weight of .50 was established as the watershed between the weights that enhance the likelihood of a certain variant's occurrence (above .50) and those that inhibit its appearance (below .50). A sample of the binomial 1-level results for the Spanish data is presented in Figure 4.

Figure 4. Binomial 1-level results for factor group **sC sonority**

Binomial Varbrul, 1 step

Name of cell file: Untitled.cel

Using fast, less accurate method.
Averaging by weighting factors.

- One-level analysis only:One-level binomial analysis:

Run # 1, 194 cells:
No Convergence at Iteration 20
Input 0.320

Group	Factor	Weight	App/Total	Input&Weight
1:	N	0.418	0.33	0.25
	S	0.540	0.41	0.36
	L	0.524	0.40	0.34

Note that the factors 'following stop (S)' and 'following liquid (L)' favor the application of [e]-epenthesis before sC clusters (.54 and .52 respectively), while the factor 'following nasal (N)' does not have a significant effect on the phenomenon (.42).

The binomial up and down analysis enables the researcher to look for significant independent variables or factor groups. The results of this kind of operation provide a summary of the groups eliminated and the best stepping up and stepping down runs. The best stepping up and stepping down runs should select and discard the same factor groups (see forthcoming discussions in 4.2.1). The selected factor groups are those that have a significant effect on the application of [e]-epenthesis.

The following section describes the instances where problematic factors or factor groups were either removed or regrouped in the different analyses conducted in this investigation.

4.2 The Goldvarb analyses

The first Goldvarb analysis included all factors originally conceived and based on the hypotheses stated in Chapter 2. However, it was observed that the analysis contained categorical results from one of the participants (i.e., no instances of [e]-epenthesis). The data elicited from this advanced EFL learner were then removed from subsequent analyses. In addition to this, results from the first statistical analysis showed that all preceding consonants (i.e., nasals, liquids, fricatives, and stops) were relatively of equal significance (i.e., they were assigned a value around .5). Therefore, they were regrouped as one single factor 'consonants', as opposed to vowels and pauses, in the factor group *preceding environments*. The results obtained from a second analysis led to the identification of a redundancy between the factor groups: *proficiency* and *subjects*. Because redundancies can lead to interference, the factor group *subjects* was removed from subsequent Goldvarb analyses. Finally, a third analysis involved the recoding of the factors 'pauses' and 'consonants' into a single factor (consonant/pause) because they both showed relatively similar significant effects on the application of [e]-epenthesis (i.e., a value of above .5). Let us now proceed to present the final Goldvarb results.

4.2.1 Final Goldvarb results

In order to identify the most significant factor groups responsible for the variable application of [e]-epenthesis, the present study carried out a binomial up and down statistical analysis. The results obtained from this analysis are presented in Figure 5 below, where the best stepping up (#9) and stepping down (#16) runs are shown. Observe that both runs selected the factor groups 1, 2 and 3 as the most significant in the phenomenon under investigation (i.e., *sC sonority*, *preceding environment*, and *proficiency*, respectively). Additionally, both runs discarded factor group 4 (i.e., *level of formality*) because it did not significantly influence the phenomenon.

Figure 5. Best stepping up and stepping down runs in MS English

```
Binomial Varbrul
================
Using fast, less accurate method.
Averaging by weighting factors.
Threshold, step-up/down: 0.050001
All remaining groups significant

Groups eliminated while stepping down:  4

Best stepping up run:      #9
Best stepping down run:    #16

Run # 9, 18 cells:
Convergence at Iteration 5
Input 0.365
Group # 1 -- N: 0.435, S: 0.534, L: 0.511
Group # 2 -- c: 0.554, v: 0.381
Group # 3 -- a: 0.295, i: 0.430, b: 0.751
Log likelihood = -886.368 Significance = 0.010

Run # 16, 18 cells:
Convergence at Iteration 5
Input 0.365
Group # 1 -- N: 0.435, S: 0.534, L: 0.511
Group # 2 -- c: 0.554, v: 0.381
Group # 3 -- a: 0.295, i: 0.430, b: 0.751
Log likelihood = -886.368 Significance = 0.113
```

In sum, the results from the binomial up and down statistical analysis indicate that the internal variables *sC sonority* and *preceding environment*, and the external variable *proficiency* have significant conditioning effects on the variable application of [e]-epenthesis in the speech of Hispanophone EFL learners. This kind of analysis, however, provides us with no information as to what individual factors in each group significantly influence the phenomenon. This information was obtained via the binomial 1-level statistical analysis, discussed in 4.1 above.

The results from the binomial 1-level statistical analysis (see Table 8 below) indicate that [e]-epenthesis is more likely to occur: before s+liquid (.52) and s+ stop (.54) onset clusters, when sC onset clusters are preceded by consonants or pauses (.57), and in the speech of less proficient speakers (.71). On the other hand, the results from the factor group *level of formality* show that the two-level distinction (i.e., formal and informal) is not significant, as indicated in the stepping up and stepping down analysis discussed in section 4.2.1, and based on the weights assigned to these two factors in the binomial 1-level analysis shown in Table 8.

Table 8. Final Goldvarb probabilistic results

Likelihood of [e]-epenthesis occurrence

Factor Groups	**Factors**		
sC sonority	s+nasal .42	s+liquid **.52**	s+stop **.54**
Preceding environment	Consonant/Pause **.57**	Vowel .34	
Proficiency level	Beginning **.71**	Intermediate .44	Advanced .32
Level of formality	Formal .49	Informal .58	

The following sections provide a more detailed discussion of the probabilistic results based on the linguistic and extralinguistic factors included in the investigation of [e]-epenthesis.

4.2.2 Results and discussion of the linguistic factors

The present study initially predicted that, in contrast to s+stop sequences, s+nasals and s+liquids would surface earlier in the process of L2 acquisition as a consequence of markedness. Specifically, this study expected marked structures (i.e., s+stop onset clusters) to be more difficult to acquire than less marked structures (i.e., s+liquid and s+nasal clusters) because the former violate sonority. The results show a pattern that is partially inconsistent with the initial hypothesis: as expected, [e]-epenthesis is more likely to occur before s+stop onset clusters; however, s+liquids also exhibited a surprisingly significant effect on the application of [e]-epenthesis. These results are illustrated in Table 9 below (From Table 8, repeated here for ease of exposition).

Table 9. Final probabilistic results for the factor group *sC sonority*

Factor group	Factors	[e] epenthesis
sC sonority	s+liquid	**.52**
	s+nasal	.42
	s+stop	**.54**

If s+liquid onset sequences are less marked than s+stop clusters, how can one account for the fact that they seem to be equally difficult for these Hispanophone EFL learners? In order to provide an explanation, let us first contemplate the idea that such a finding is not exclusively due to violations of sonority. It might represent further evidence for Carlisle's

(1991a) claim that if epenthesis occurs more frequently after consonants before /sl/, then the preceding environment is a more powerful factor in inducing epenthesis than is the sonority relationship among the members of the onset. To explore such an idea, it is necessary first to provide the results from the second linguistic factor group *preceding environment* and then proceed to discuss the results from the cross tabulations between *preceding environment* and *sC sonority*.

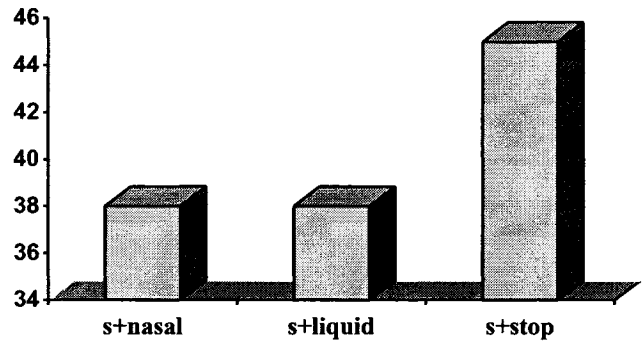
The results for the second linguistic factor, *preceding environment*, were not consistent with the initial hypothesis since the statistical program rendered significant both preceding consonants and preceding pauses (see Table 10 below). Recall from section 2.6 that, based on Abrahamsson's (1999) study in which pauses had no effect on the frequency of epenthesis, it was predicted that consonants would induce the highest frequency of epenthesis. The results obtained in this study, however, show that pauses behave like consonants in the sense that they also induce a high frequency of epenthesis (i.e., when compared to preceding vowels). The fact that preceding pauses and consonants behave in a similar way is not surprising – this is in fact consistent with a number of variationist studies (e.g., Winford, 1992; cf. Cardoso, 1999). Specifically, preceding consonants and pauses hinder the production of sC clusters because the first member of the cluster cannot be resyllabified as the coda of the preceding syllable. In contrast, preceding vowels allow Hispanophone learners to break the sC cluster because they reduce the necessity of applying [e]-epenthesis.

Table 10. Final probabilistic results for the factor group *preceding environment*

Factor groups	Factors	[e] epenthesis
Preceding environment	Consonant	**.59**
	Vowel	.34
	Pause	**.55**

Observe that the results above express the influence of preceding environments independently of other factor groups (e.g., sC sonority, proficiency, formality). The influence of preceding environments on the application of [e]-epenthesis across types of sC clusters can be obtained via a cross-tabulation. Recall from section 4.1 that a cross-tabulation allows one to check for interactions between factor groups and identify the presence of categorical results. The cross-tabulation between *preceding environment* and *sC cluster* revealed that among all preceding environments (i.e. consonants, pauses, and vowels), preceding pauses exhibited a pattern of [e]-epenthesis application that seems to abide to the marked – unmarked dichotomy, as illustrated in Figure 6.

Figure 6. [e]-epenthesis by pause and sC cluster (%)



Observe that when the sC onset clusters occurred after pauses (i.e., no preceding consonants or vowels), liquids and nasals showed a relatively similar percentage of

epenthesis. These results suggest that when all sC clusters were preceded by a pause, sonority played a significant role on the production of these clusters by Hispanophone EFL learners. Specifically, the less marked onset clusters were produced more successfully than the more marked sequences. This is consistent with the initial hypothesis that the less marked s+liquid and s+nasal clusters would surface as such before the more marked s+stop.

The discussions above suggest that no satisfactory explanation can be reached if one tries to explain the variable production of sC clusters with one single linguistic factor in mind (e.g., sonority only). Carlisle's (1991a) claim that preceding environment is a more powerful factor in inducing epenthesis if epenthesis occurs more frequently after consonants before /sl/ appears to be relevant for the present study: the variable production of sC clusters is noticeably influenced by a combination of at least these two linguistic factors.

Let us now proceed to discuss the results from the cross-tabulations between *sC sonority* and *preceding environment* to determine which factor group is more powerful in the application of [e]-epenthesis, based on Carlisle's (1991a) proposal. According to the author, to find out if preceding environment is a more significant factor than sonority, it is necessary to explore the interaction between these factors. Specifically, if s+liquids show a high frequency of epenthesis when preceded by consonants, then it is possible to claim that preceding environment is the main factor influencing epenthesis. Conversely, if s+liquids exhibit a low frequency of epenthesis even when preceded by consonants, then it is possible to state that sonority is a more powerful factor. The cross-tabulations between *sC sonority* and *preceding environment* exhibit a high frequency of [e]-

epenthesis in consonants preceding s+liquids (44%)[7]. This finding seems to confirm the hypothesis that, when producing /sl/ onset clusters, *preceding environment* is a more powerful factor for the group of EFL learners included in this study.

In general, the unexpected results could also be due to the large amount of collected tokens coming from the formal task (i.e., 89%) as opposed to those from the informal task (i.e., 11%). Such difference is perhaps due to the fact that, while the number and types of sC clusters and preceding consonantal environments can be controlled in formal tasks, these variables cannot be easily controlled in spontaneous speech. In fact, advanced learners produced the majority of the tokens collected during the informal task (5%), while both beginning and intermediate learners together produced approximately 6%. When producing spontaneous speech, beginning and intermediate L2 learners could have made use of communication strategies such as avoidance and synonyms to express their ideas (e.g., avoiding the production of sC clusters or by saying 'begin' instead of 'start'). The use of these hypothetical strategies might have exerted an influence on the number of tokes containing sC clusters.

In addition, it appears that the adoption of Lin's (2003) grammaticality judgment task in this investigation elicited informal rather than formal speech because the explicit focus on grammar took away participants' focus on pronunciation and possibly influenced the application of [e]-epenthesis across types of sC clusters.

Finally, the amount of exposure to sC clusters in the L2 classroom might have an effect on the application of [e]-epenthesis. Specifically, an L2 classroom that provides

[7] Consonants preceding s+stops exhibited a 47% of [e]-epenthesis application. This finding is not surprising because s+stops are the most marked sC clusters, and because preceding consonantal environments hinder the production of sC clusters. s+nasals exhibited the lowest percentage of [e]-epenthesis (i.e., across sC clusters) when preceded by consonants (37%).

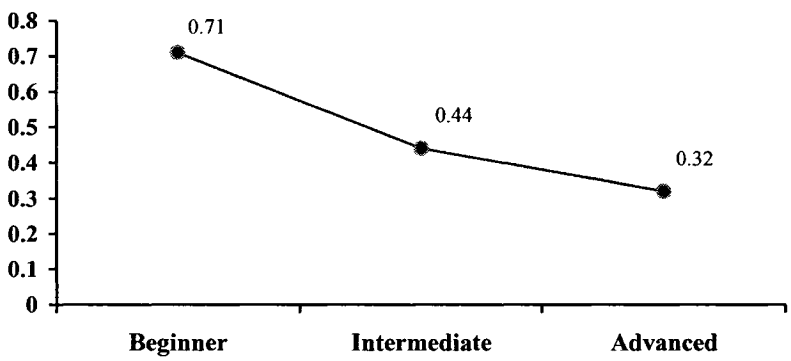
learners with frequent exposure to sC clusters via a textbook or pronunciation activities might significantly influence the production of these clusters. A verification of word frequency in the L2 as well as audio recordings of the sC clusters learners have been exposed to at a particular language institute/school might shed some light on how the L1 transfer patterns (i.e., [e]-epenthesis) are affected by these factors.

The following section presents and discusses the results for the extralinguistic factors included in this study.

4.2.3 Results and discussion of the extralinguistic factors

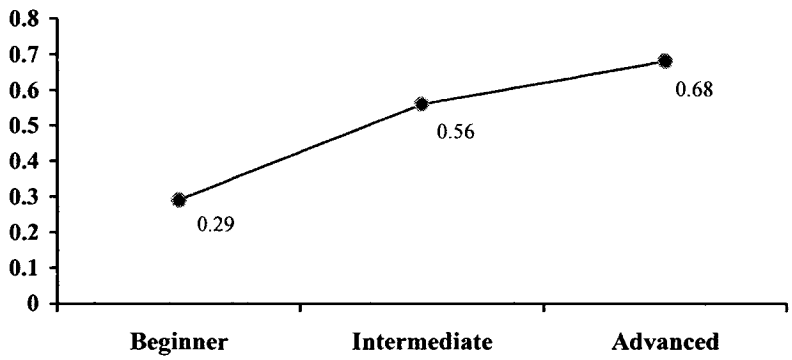
The external variable *proficiency* exhibited a significant effect on the application of [e]-epenthesis. Recall from section 2.6 that the present study predicted that the likelihood of [e]-epenthesis would decrease with increased proficiency. The results obtained from the final Goldvarb analysis show a decreasing pattern from high to low probabilities across the three proficiency groups as illustrated in Figure 7.

Figure 7. Decreasing [e]-epenthesis across proficiency levels



Observe that there is a decreasing pattern of [e]-epenthesis from .71 in the beginner group to .32 in the advanced group. These findings are consistent with the predictions made by the Ontogeny Phylogeny Model (Major, 2001). As discussed in Chapter 2, The Ontogeny Phylogeny Model (OPM) claims that over time and as style becomes more formal, L2 features increase while those from the L1 decrease. Figure 8 shows the increasing pattern of sC clusters across the three proficiency levels included in this study.

Figure 8. Increasing sC clusters across proficiency levels



The findings in Figures 7 and 8 are consistent with previous studies (e.g., Bunta & Major, 2004; Cardoso, 2004; John, to appear) which show that, as learners' pronunciation becomes more nativelike, L1 transfer substitutions diminish. It should be noted however, that the results provided above do not shed light on the influence of proficiency across sC clusters and preceding phonological environments. This information was obtained via two cross-tabulations: [e]-epenthesis by *proficiency* and *preceding environment*, and [e]-epenthesis by *proficiency* and *sC cluster*. Figures 9 and 10 illustrate the percentages from these cross-tabulations.

Figure 9. [e]-epenthesis by sC cluster and proficiency

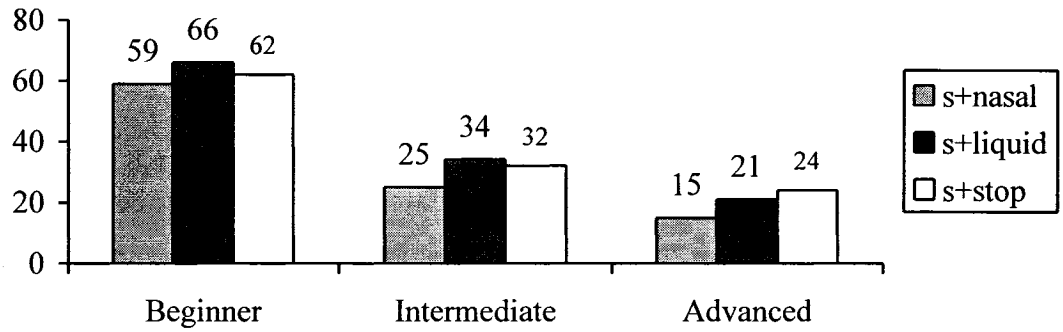
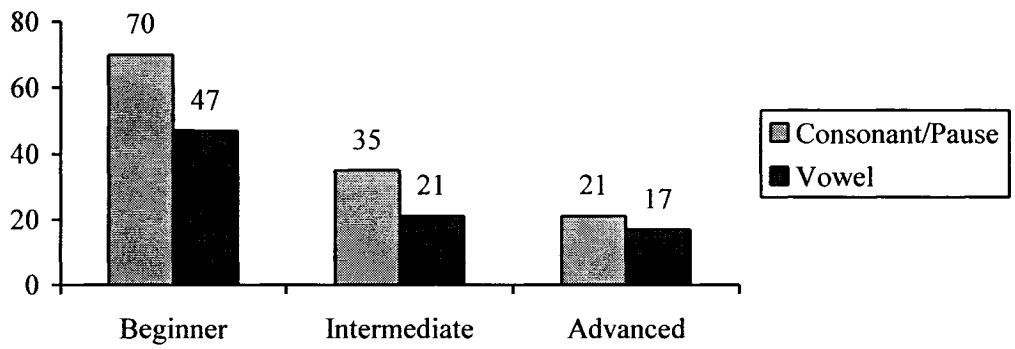


Figure 10. [e]-epenthesis by proficiency and preceding phonological environment



Observe in Figure 9 that the percentage of [e]-epenthesis across sC clusters is the highest in the group of beginning learners and that this percentage decreases gradually as learners become more proficient in the L2. Similarly, Figure 10 shows that the percentage of [e]-epenthesis after all preceding environments (i.e., consonant/pauses and vowels) decreases with increased L2 proficiency.

The results discussed in this section demonstrated that *proficiency* is unquestionably a factor that influences the application of [e]-epenthesis before sC

clusters. It should be pointed out, however, that it is relevant to investigate how the two formality levels adopted in this investigation influence the frequency of [e]-epenthesis across proficiency levels. This information was obtained via a cross-tabulation between the factor groups *proficiency* and *level of formality*. Before discussing such results, let us first present the statistical results for the extralinguistic factor *level of formality*.

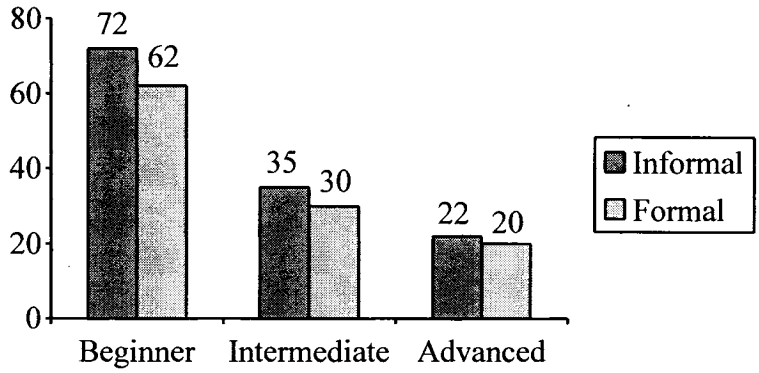
The results obtained for the second extralinguistic factor *level of formality* (see Table 11) indicate that the informal task induced a slightly higher likelihood of [e]-epenthesis occurrence (.58) in comparison with the more formal task (.49). The Goldvarb statistical analysis rendered the difference between formality levels insignificant. More precisely, results suggest that because both levels of formality reached significance on the likelihood of [e]-epenthesis occurrence, Goldvarb was not able to determine which factor had a more significant effect.

Table 11. Final probabilistic results for the factor group *formality level*

Factor group	Factors	[e] epenthesis
Formality level	Formal	.49
	Informal	.58

Clearly, the results from the statistical analysis illustrated above provide no information as to how the frequency of [e]-epenthesis varies overtime and across formality levels. The cross-tabulation illustrated in Figure 11 examines the interaction between these factor groups, in percentage.

Figure 11. [e]-epenthesis by proficiency and formality (%)



The present study expected to find stylistic distinctions for intermediate and advanced learners (i.e., a formal and an informal grammar), possibly no stylistic distinctions for beginning learners[8], and a higher frequency of [e]-epenthesis during the formal task because a number of sociolinguistic studies indicate that L1 transfer decreases in more formal tasks because L2 speakers monitor their speech more closely (Dickerson, & Dickerson, 1977; Gatbonton, 1978; Major, 2001; Major, 2004; Cardoso, 2004; John, to appear). The results in Figure 11, however, indicate that while beginning learners exhibit a slightly higher (albeit insignificant) level of stylistic distinctions (i.e., 62% in the formal task and 72% in the informal task), intermediate and advanced learners show very little difference between the formal and informal tasks.

A possible explanation for the unexpected findings involving the group of beginners may be the fact that, in the less formal task, learners' speech could have been affected by extralinguistic factors such as nervousness and task difficulty. More

[8] This assumption is based on the idea that the grammar of beginning learners is characterized by monostylism (Cardoso, 2004). Such monostylism is possibly due to the limited exposure to different stylistic varieties in earlier stages of the L2, as is usually the case in L1 acquisition.

precisely, learners' speech could have been affected by the exposure to a difficult task such as the informal interview. This unfamiliar task produced a feeling of uneasiness or nervousness as overtly expressed by one of the participants at the end of the data collection session. Alternatively, intermediate and advanced learners showed no differences because the language institute where the data collection took place adopts a communicative approach to language teaching that not only influenced participants' familiarity with informal interviews, but also provided them with sufficient practice to produce more careful speech in such contexts.

As mentioned in section 4.2.2, it is possible that Lin's (2003) grammaticality judgment task adopted in this investigation influenced the application of [e]-epenthesis across sC clusters because of its focus on grammar. An increased percentage of [e]-epenthesis during this task would therefore lead to a less clear distinction between the two tasks incorporated in this study. Possibly, the incorporation of a more formal task such as the reading of wordlists will allow the collection of more formal data.

In view of the fact that the study did identify differences in the likelihood of [e]-epenthesis occurrence across proficiency levels (see Figure 7), this thesis adopts the standard view that each proficiency level represents an interlanguage, and consequently, a grammar (e.g., Selinker, 1972; Adamson, 1988; Preston, 1996; Cardoso, 2004; John, to appear).

In sum, this section has shown that variation in the acquisition of sC onset clusters by Hispanophone EFL learners is conditioned by a combination of linguistic (i.e., *sC sonority* and *preceding phonological environment*) and extralinguistic factors (i.e., *proficiency*). Specifically, the study has shown that sC onset clusters are more likely to

occur in s+nasal clusters, when sC clusters are preceded by a vowel, and in the speech of more proficient speakers.

The following chapter shows how some of the tools provided by current developments in phonological theory can account for the variable results obtained in this study.

5. STOCHASTIC OT

This chapter provides an analysis for the variable results obtained in the study within a stochastic version of Optimality Theory. It is argued that the variation encountered in the corpus is motivated by the interaction of a set of constraints that overlap in their normal distribution.

5.1 Introduction to Optimality Theory

According to Optimality Theory (Prince & Smolensky, 1993), phonological systems are the result of rankings of universal constraints, which are part of the grammars of all natural languages. The framework of OT recognizes two types of constraints: markedness and faithfulness.

Markedness constraints require that output forms meet some criteria of structural well-formedness. For instance, the constraint ***sC** proposed in this study captures the cross-linguistic observation that s+consonant clusters are disallowed in several languages (e.g., Spanish, Arabic, Portuguese, Bengali), and are only acquired later in L1 acquisition (e.g., Goad & Rose, 2004). *sC is a cover term for a family of constraints that ban s+consonant clusters in onset position. In the case of Spanish, when the language borrows English words such as /**sl**ɪm/, sC clusters are produced with the epenthetic vowel [e] as in [**es.l**ɪm]. Note that languages that allow sC clusters (e.g., English, French), incur in violations of this constraint.

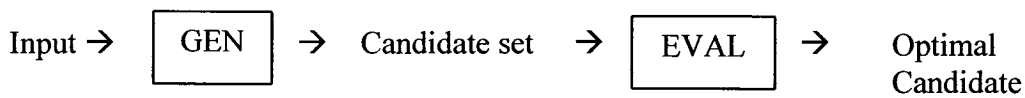
The second type of constraints, faithfulness, requires that outputs preserve the properties of their basic (lexical) forms, requiring some kind of similarity between the output and its input. For example, the constraint **DEP-IO** states that output segments

must have input correspondents (i.e., no epenthesis). The effect of this constraint can be easily detected in languages such as English and Swedish, where sC clusters surface intact. In these languages, an output like *[estɛp] 'step' is disallowed because the epenthetic vowel [e] does not have a correspondent in the input /stɛp/. Let us now address the basic operations involved in the selection of an optimal output in OT.

5.1.1 The basic operations of OT

In OT, the selection of an optimal output is accomplished through a series of basic operations. First, the input (i.e., the underlying form) goes into the function Generator (GEN), which creates an infinite number of possible candidates or outputs. The evaluator (EVAL) evaluates all possible outputs based on the language-specific rankings of the constraint inventory (Con) provided by UG, and selects the most optimal candidate (i.e., the form that incurs in fewer or minimal violations of highly ranked constraints). A schematic representation of these operations is illustrated in (12).

(12) The basic operations of OT



A constraint inventory (Con) contains universal constraints such as those that reflect markedness and faithfulness. These constraints are ranked on a language particular basis

to regulate conflicts between structural well-formedness (markedness) and the preservation of lexical forms (faithfulness). Let us illustrate the basic operations of OT with a concrete example. Suppose that Con consists of DEP-IO (no epenthesis), *sC (no sC clusters), and MAX-IO - a constraint that requires input segments to have output correspondents (i.e., no deletion), and that English is defined by the ranking illustrated in (13).

(13) Constraint ranking in English: MAX-IO, DEP-IO >> *sC

Observe that the constraints on the left (i.e., MAX-IO and DEP-IO) are the highest ranked and that the constraint on the right (i.e.,*sC) is the lowest ranked. Double arrowheads indicate that the constraints are crucially ranked with respect to one another, while commas indicate that the ranking of the constraints (e.g., MAX-IO and DEP-IO in (13)) is indeterminate. The evaluation of a ranking such as the one provided above can be illustrated via a tableau.

By OT convention, a tableau illustrates a language-specific ranking of constraints, the possible candidates, and their evaluation. A solid line in the tableau indicates that two constraints are crucially ranked with respect to each other, while a dotted line indicates that the ranking is indeterminate. The winning candidate is indicated with a hand on the left. An asterisk (*) indicates a violation of one of the constraints, and an exclamation mark after an asterisk (*!) indicates a fatal violation, which signals that the candidate will not surface. The cells for the lower ranked constraints are shaded to emphasize the

irrelevance of these constraints in the selection of the optimal candidate. Let us now proceed to evaluate the constraint ranking in (13) via the tableau illustrated below.

Tableau 1. Constraint evaluation in English

/stɛp/	MAX-IO	DEP-IO	*sC
a.[estɛp]		*!	
☞ b.[stɛp]			*
c.[tɛp]	*!		

Note in the tableau that some conceivable outputs (i.e., (a), (b), and (c)) violate at least one of the constraints in the inventory. In OT, constraint violation does not lead to ungrammaticality. The most optimal output is the one that *minimally* violates a constraint. Candidate (a) [estɛp] was ruled out as optimal because it incurs in a fatal violation of the highly ranked constraint DEP-IO: the epenthetic [e], which is not in the input, is present in the output. Candidate (b) [stɛp] was selected as optimal even though it violates the constraint *sC. In English, minimal violations to *sC are allowed as long as higher ranked constraints (i.e., MAX-IO and DEP-IO) are satisfied. Finally, candidate (c) [tɛp] was rejected as optimal because it fatally violates the constraint MAX-IO (i.e., due to the deletion of the input segment /s/ from the output). Notice that the ranking of MAX-IO and DEP-IO is indeterminate. This suggests that in English, it is equally fatal to violate either of these constraints, and it is indicative of the tendency in the language to satisfy faithfulness constraints.

As mentioned earlier, constraints are ranked on a language-particular basis and, thus, differences in constraint rankings define different grammars or languages. To illustrate cross-linguistic variation, suppose that, as was the case for English, Con for Spanish consists of MAX-IO, DEP-IO, and *sC, and that this language is defined by the constraint ranking in (14).

(14) Constraint ranking in Spanish: MAX-IO, *sC >> DEP-IO

Observe that in Spanish the constraint *sC is ranked above DEP-IO. Furthermore, MAX-IO and *sC are ranked equally high with respect to one another. The evaluation of this constraint ranking is illustrated in Tableau 2.

Tableau 2. Constraint evaluation in Spanish

/stɛp/	MAX-IO	*sC	DEP-IO
☞ a.[estɛp]			*
b.[stɛp]		*!	
c.[tɛp]	*!		

Notice that candidate (a) [estɛp] was selected as the most optimal candidate even though it is not the most faithful to the input (i.e., it violates DEP-IO). In Spanish, minimal violations to DEP-IO are allowed because both MAX-IO and *sC must be satisfied (i.e., they are ranked at the higher end of the hierarchy). Candidates (b) [stɛp] and (c) [tɛp]

were discarded as optimal because they incur in fatal violations of the constraints *sC and MAX-IO respectively. More precisely, deletions and sC clusters constitute more serious violations than [e]-epenthesis in a language like Spanish.

In sum, this section has shown that OT allows one to account for variation across languages. The following section demonstrates how this framework is also able to account for variation that takes place within a single language or language variety (e.g., within an advanced interlanguage grammar).

5.2 Variation in OT

In standard OT, constraints are strictly ranked (e.g., *sC >> DEP-IO) and the strict ranking of constraints is responsible for the selection of a single optimal output. The question that remains to be answered is: how can variation be explained within this approach? More precisely, how can one account for instances where more than one optimal output is observed? The present section discusses three approaches that have been proposed for the analysis of variation within the framework of OT (i.e., the multiple grammars approach, crucial non-ranking of constraints, and stochastic OT). Let us start by providing an introduction to the first two approaches, which is then followed by a discussion of their abilities to predict output frequencies. The section ends with the introduction and discussion of Stochastic OT, the approach adopted in this study.

The first approach is the *Multiple grammars approach* (Kiparsky, 1993), where variation results from the competition of different grammatical systems within an individual (i.e., an individual commands a set of different grammars or constraint rankings). When producing a word, for instance, the speaker reaches into the grammar

pool and selects a ranking. For example, the grammar of a Hispanophone EFL learner can consist of the two rankings provided in (15) below. Observe that while grammar (15a) selects [e]-epenthesis, grammar (15b) produces the most faithful, target-like output.

(15) A Spanish-English grammar

Input	Ranking	Output
a. /stɛp/	MAX-IO >> *sC >>DEP-IO	[estɛp]
b. /stɛp/	MAX-IO >> DEP-IO >> *sC	[stɛp]

There are however, a few problems with this approach. First, the number of grammars per individual can become rather large because any set of rankings represents a plausible grammar. Second, the approach poses conceptual problems for L2 acquisition. More precisely, learners have to acquire a large number of rankings in the process of developing an L2, which goes against one of the principles of linguistic theory, the principle of parsimony or Occam's razor. Let us now consider how a different approach can account for variation and output frequencies.

Following Reynolds' (1994) floating constraints approach[9], Anttila (1997) proposed a model where constraints are partially ranked via crucial non-ranking of constraints. Anttila's (1997) approach, which is also a development of Kiparsky's (1993) approach, assumes that a constraint set can impose crucial non-dominance (i.e., crucial nonranking) of its components. The result is the possibility of two or more acceptable

[9] According to this approach, a grammar is defined by a single constraint hierarchy where one or more constraints may float with respect to another constraint or set of constraints. Because Reynolds' view on variation overlaps with that of Anttila, the focus in this discussion will be on the latter. For a detailed review of this approach, see Cardoso (2003).

outputs in that grammar, which leads to variation. Within this model, the probability of variant's occurrence is predicted in the following way:

(16) Anttila's variant probabilistic prediction

a) A candidate is predicted by the grammar iff it wins in some tableau

b) If a candidate wins in *n* tableaux and *t* is the total number of tableaux, then the candidate's probability of occurrence is *n/t*. (Anttila, 1997, p.40).

To illustrate this, assume that in the grammar of Hispanophone EFL learners (SE henceforth), the constraints *sC and DEP-IO are partially ordered as exemplified in (17) below. The semicolon (;) indicates the crucial non-ranking of these constraints with respect to each other, while the curly brackets ({}) separate the set of crucially unranked constraints.

(17) A partially ranked grammar: MAX-IO » {***sC ; DEP-IO**}

Following Anttila's variant probabilistic prediction in (16), two possible constraint rankings derive from the grammar exemplified above. One in which *sC outranks DEP-IO, and another in which DEP-IO outranks *sC. This is illustrated in (18).

(18) Possibilities of rankings

MAX-IO » ***sC** » DEP-IO

MAX-IO » DEP-IO » ***sC**

The two constraint rankings above indicate that two optimal forms are possible in a given SE grammar (e.g., advanced). More precisely, the candidate [estɛp] is selected when *sC is ranked higher than DEP-IO, while the candidate [stɛp] is selected in the reverse situation. The tableaux below exemplify this.

Tableau 3. Variation in SE

Tableau (a)
MAX-IO » *sC » DEP-IO

/stɛp/	MAX-IO	*sC	DEP-IO
☞ a.[estɛp]			*
b.[stɛp]		*!	
c.[tɛp]	*!		

Tableau (b)
MAX-IO » DEP-IO » *sC

/stɛp/	MAX-IO	DEP-IO	*sC
a.[estɛp]		*!	
☞ b.[stɛp]			*
c.[tɛp]	*!		

So far, it has been shown that this approach is able to account for variation within one single grammar via the crucial non-ranking of the constraints *sC and DEP-IO. We should now address the issue of how the approach accounts for output frequencies.

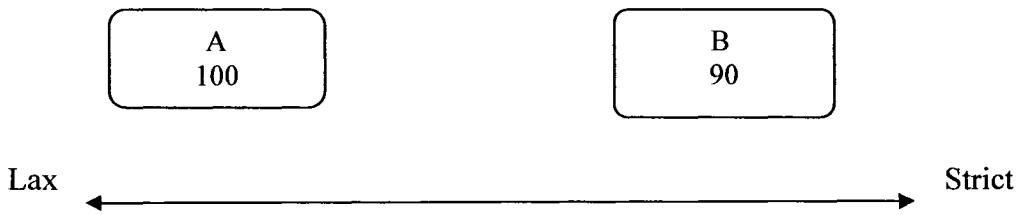
According to Anttila's (1997) variant probabilistic prediction (see (16) above), candidates (a) and (b) in Tableau 3 win in one tableau each (n = 1), and the total number of tableaux is two (t = 2). n/t = ½ = 0.5 or 50%. Each candidate's probability of occurrence is 0.5 and each variant is likely to occur 50% of the time in each grammar.

One of the problems of the crucial non-ranking of constraints and the multiple grammars approach is that they are not able to accurately capture the observed frequencies in the data under investigation. For instance, the group of beginning SE learners investigated in this study, exhibited a .71 likelihood of [e]-epenthesis application, and a .29 likelihood of sC occurrence. These findings clearly contradict the predicted 50-50 per cent probability of sC and [e]-epenthesis occurrence predicted by the crucial non-ranking of constraints approach. A tentative solution would be to increase the number of constraints in the inventory. This solution, however, is not parsimonious if the number of constraints in the inventory is increased solely with the purpose of obtaining the right frequencies.

Another approach to investigating and representing variability in Optimality Theory is *Stochastic OT* (Boersma & Hayes, 2001). This approach includes a Gradual Learning Algorithm (GLA henceforth), which is a development of Tesar and Smolensky's (1995, 1998) Constraint Demotion algorithm. Unlike standard OT, SOT adopts a continuous ranking scale, where each constraint has a fixed ranking value along a real number scale and higher values correspond to higher-ranked constraints. The

ranking value is called the selection point. The GLA assumes that selection points for constraints are distributed normally with the mean of the distribution of the ranking value (i.e., normal distributions have the same standard deviation for every constraint: 2.0). Within this approach, variation is established through the distance between constraints on a strictness scale, and by the amount of evaluation noise (i.e., the standard deviation of the distribution) added to the strictness scale. More precisely, ranking values determine whether a grammar is categorical or variable, and encode into the grammar the number of times a constraint will overrank another (i.e., it captures probabilities). Figure 12 below depicts a categorical grammar, where two ranked constraints are distant from each other. Observe that A is ranked 10 points higher than B. Therefore, this grammar will always rank A at the higher end of the hierarchy, and only one categorical output can be selected. This is the equivalent of A >> B in standard OT.

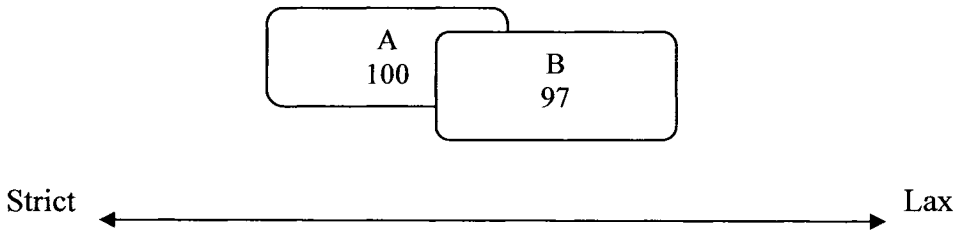
Figure 12. A categorical grammar



This leaves us with the question of how a categorical grammar differs from a variable one. In a variable grammar, the distribution of crucially ranked constraints overlaps because their ranking values are close to one another, as illustrated in Figure 13 below. This grammar might select any point within the overlap of constraint A and B.

Because A is ranked above B, the grammar is likely to select A >> B more frequently. However, the grammar can also select a point within the higher ranked area of B and the lower ranked area of A. When this occurs, B will be ranked higher than A and a different candidate (i.e., output) will be selected.

Figure 13. A variable grammar



For the analysis of variation in SE, the main advantages of SOT are: (1) we are able to account for variation within one single grammar, and (2) probability distributions can be accurately described and encoded into the grammar. Anttila's (1997) approach predicts that quantities of variation should be in smaller integer fractions (e.g., 2/3, 1/2, 1/3). As shown in Chapter 4, the data in this study do not show this type of variation. The following section presents the Stochastic OT analysis for the variable results presented in Chapter 4.

5.3 The Stochastic OT analysis

The present study adopts three OT constraints for the stochastic analysis of the variable data, which are repeated below from section 5.1.1 for ease of exposition.

(19) Relevant OT constraints

MAX-IO	Input segments must have output correspondents (no deletion) (McCarthy and Prince, 1995)
DEP-IO	Output segments must have input correspondents (no epenthesis) (McCarthy and Prince, 1995)
*sC	No s+consonant clusters in onset position

These constraints along with a set of inputs, candidates and violations were employed to prepare an input file for every grammar in the data set (see forthcoming discussion about the grammars adopted in this investigation). The quantitative values (weight) established by Goldvarb were fed into the software package OTsoft (Hayes et al 2003), which contains a Gradual Learning Algorithm (GLA) that simulates the learning process and assigns constraints ranking values.

OTsoft individually learned the grammars that represent the data set by exposing its algorithm to one million input data (evaluation noise: 2.00, initial/final plasticity: 2/00.2, original arbitrary ranking for each constraint: 100). At the end of the simulations, the algorithm arrived at a final grammar, which provides the relative frequency of variants in the data by assigning a ranking value for each of the abovementioned constraints. The next section presents a discussion of the grammars that represent the data set, followed by the stochastic analysis of each grammar involved in the study.

5.3.1 The SE grammars

As you will recall from section 4.2.3, the probabilistic results for proficiency revealed a three-level distinction in the likelihood of [e]-epenthesis across the three levels of proficiency included in this investigation. It was also assumed, based on standard literature on the subject, that each proficiency level represents an interlanguage and, by definition, a separate grammar (e.g., Selinker, 1972; Adamson, 1988; Preston, 1996; Cardoso, 2004): Beginners, Intermediate, and Advanced.

We will now re-examine the statistical results obtained for each of these three proficiency levels and subsequently provide a stochastic OT analysis that will account for the variable development of sC clusters across these grammars. Let us start with the grammar that represents that of the group of Beginners.

The probabilistic results illustrated in section 4.2.1 indicate that, in the grammar that characterizes the group of beginning learners, the likelihood of sC occurrence is .29, while the probability of [e]-epenthesis to occur is .71. These probabilities were learned by the GLA in order to generate the ranking values illustrated in Table 12.

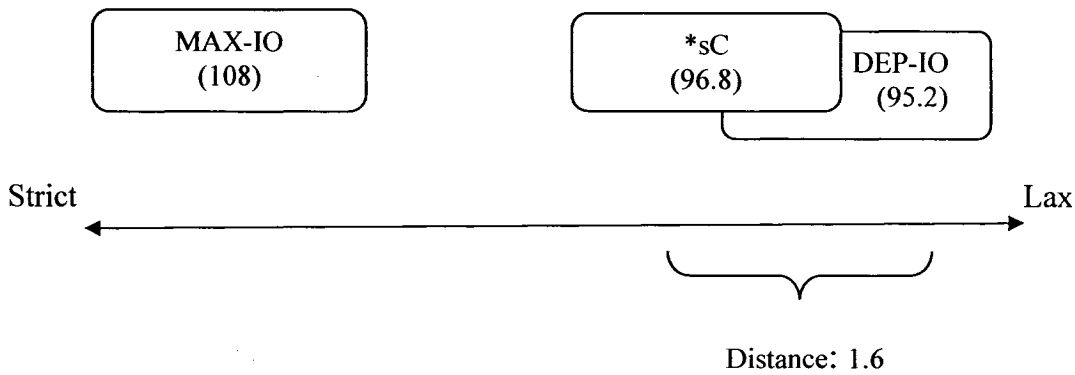
Table 12. Beginners' grammar: ranking values

Constraint	Ranking value
MAX-IO	108
*sC	96.8
DEP-IO	95.2

The shaded constraints *sC and DEP-IO indicate that these constraints overlap in their distribution and consequently result in variation. For ease of exposition, Figure 14 below

shows a graphical representation of the constraints, their ranking values and the distance between constraints[10].

Figure 14. Beginners' grammar



Because MAX-IO is ranked 11.2 units higher than the other two constraints, this grammar will always rank MAX-IO at the higher end of the hierarchy and consequently segmental deletion will never occur (e.g., /s/ → *ø). Note in Figure 14 above that *sC and DEP-IO overlap because of the relatively close ranking values assigned to these two constraints. According to these values, *sC will outrank DEP-IO in most evaluations (i.e., 73% of the time) because of its higher ranking value, and the outcome of such ranking will be [e]-epenthesis. This constraint evaluation is illustrated in Tableau 4.

[10] Note that the Figure does not represent accurate distances between constraints; these distances are provided for illustrative purposes only.

Tableau 4. Constraint evaluation for beginning grammar: [e]-epenthesis

MAX-IO » ***sC** » **DEP-IO**

/stɛp/	MAX-IO	*sC	DEP-IO
☞ a.[estɛp]			*
b.[stɛp]		*!	
c.[tɛp]	*!		

This variable grammar might also select a point within the higher ranked area of DEP-IO and the lower ranked area of *sC. When this occurs, DEP-IO will be ranked higher than *sC (i.e., 27% of the time) and [stɛp] will be selected as the optimal output, as exemplified in Tableau 5.

Tableau 5. Constraint evaluation for beginning grammar: sC

MAX-IO » **DEP-IO** » ***sC**

/stɛp/	MAX-IO	DEP-IO	*sC
a.[estɛp]		*!	
☞ b.[stɛp]			*
c.[tɛp]	*!		

The number of rankings established by the ranking values assigned by the GLA is able to encode into the grammar the relative frequency of the outputs generated by the grammar. This is illustrated in Table 13 where the GLA predicted that *sC will outrank DEP-IO 73% of the time, and the opposite pattern will hold 27% of the time.

Table 13. Output selection for Beginners' grammar

	Output selection		Frequency (%)	
Constraint rankings	sC	[e]-epenthesis	GLA	observed
a. MAX-IO » *sC » DEP-IO		✓	.73	.71
b. MAX-IO » DEP-IO » *sC	✓		.27	.29

Note that the frequencies obtained by the GLA (under *GLA*) closely match the probabilities observed in the data (under *observed*), and reflect the high likelihood for these learners to apply [e]-epenthesis. Let us now provide the SOT analysis for the intermediate grammar.

The grammar of intermediate learners is characterized by a .44 probability of [e]-epenthesis occurrence in both the formal and informal tasks, and a .56 probability of sC occurrence. These probabilities were learned by the GLA and the ranking values illustrated in Table 14 were generated.

Table 14. Intermediate grammar: ranking values

Constraint	Ranking value
MAX-IO	106
DEP-IO	97.3
*sC	96.8

Observe that the constraint DEP-IO was assigned a value of 97.3, which indicates that this constraint outranks *sC in this grammar, and thus predicts a higher likelihood of [e]-epenthesis in the grammar of intermediate learners. The graphic representation of these constraints is shown in Figure 15.

Figure 15. Intermediate grammar

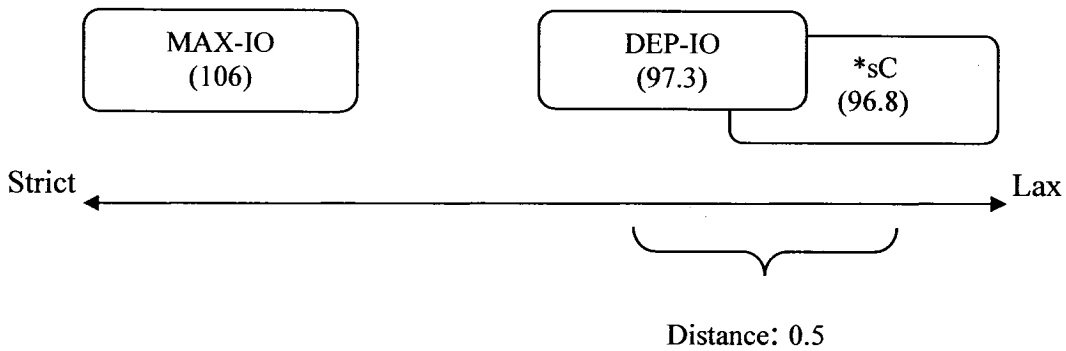


Figure 15 shows that, unlike the beginner's grammar, in the intermediate grammar, the constraint *sC is at the lower end of the hierarchy while DEP-IO is ranked higher. Because DEP-IO is ranked 0.5 units higher than *sC, their grammar is more likely to produce sC clusters (i.e., a .57 probability of occurrence). Clearly, this ranking is indicative of a higher likelihood of sC occurrence in the grammar of intermediate learners. Table 15 illustrates the frequencies obtained by the GLA, which are consistent with the probabilities observed in the corpus under investigation.

Table 15. Output selection for the Intermediate grammar

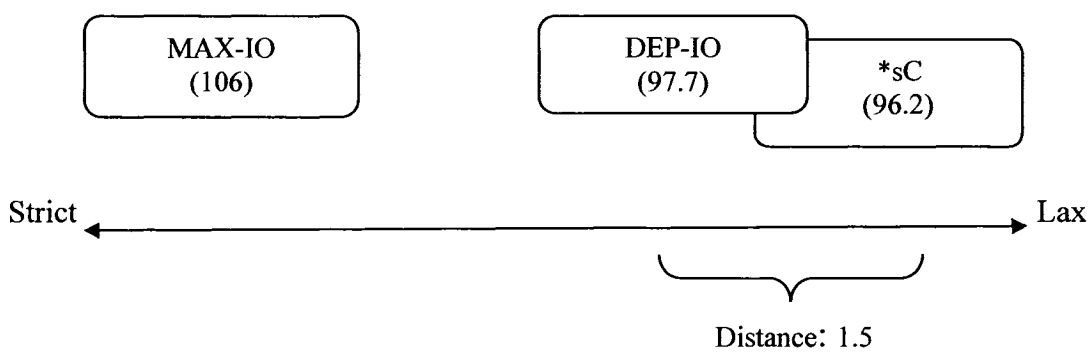
Constraint rankings	Output selection sC	Output selection e-epenthesis	Frequency (%) GLA	Frequency (%) observed
a. MAX-IO » DEP-IO » *sC	✓		.57	.56
b. MAX-IO » *sC » DEP-IO		✓	.43	.44

Let us now present the results for the advanced grammar. These advanced learners showed a .32 likelihood of [e]-epenthesis during both the formal and informal tasks and a .68 probability of sC occurrence. The ranking values assigned to the constraints are shown in Table 16, and the graphical representation of the constraints is presented in Figure 16 below.

Table 16. Advanced grammar: ranking values

Constraint	Ranking value
MAX-IO	106
DEP-IO	97.7
*sC	96.2

Figure 16. Advanced grammar



Notice that in this grammar DEP-IO has moved further away from *sC. More precisely, the distance between DEP-IO and *sC (i.e., 1.5) is greater than that found in the intermediate grammar (i.e., 0.5), which results in a higher probability of sC occurrence (.70). However, [e]-epenthesis is still likely to occur with a probability of .30. The frequency in which [e]-epenthesis and sC occur, and their respective GLA predictions are illustrated below.

Table 17. Output selection for the Advanced grammar

	Output selection		Frequency (%)	
Constraint rankings	sC	e-epenthesis	GLA	observed
a. MAX-IO » DEP-IO » *sC	✓		.70	.68
b. MAX-IO » *sC » DEP-IO		✓	.30	.32

In sum, each grammar learned by the GLA generates output frequencies similar to the probabilities observed in the data. A summary of the grammars by proficiency and style is illustrated below.

Table 18. Summary of Grammars by Proficiency and Style

Grammars by Proficiency and Style	
Beginner:	MAX-IO^{108} >> *sC$^{96.8}$ >> DEP-IO$^{95.2}$
Intermediate:	MAX-IO^{106} >> DEP-IO$^{97.3}$ >> *sC$^{96.8}$
Advanced:	MAX-IO^{106} >> DEP-IO$^{97.7}$ >> *sC$^{96.2}$

Overlapping constraints

Note in Table 18 that in the beginners' grammar, *sC outranks DEP-IO. In addition, the distance between these two overlapping constraints, as determined by their ranking values, results in a higher likelihood of [e]-epenthesis (.73). In the intermediate and advanced grammars, on the other hand, DEP-IO outranks *sC. More precisely, the ranking values that determine the distance between the overlapping constraints in the intermediate grammar (0.5) indicate that this grammar is characterized by a higher tendency to produce sC clusters (i.e., when compared to the grammar of beginning learners). Finally, the advanced grammar is characterized by an even greater distance between DEP-IO and *sC (1.5), resulting, thus, in a higher likelihood of target-like sC forms.

To conclude, stochastic OT is able to account for variation and accurately predict frequencies of different variants within one single grammar. This has important consequences for both the study of variation and linguistic theory because it allows one to

narrow down the competence – performance dichotomy. Specifically, this study assumes that competence is a much broader term than that proposed by Chomsky (1965) since "[t]he ability of human beings to accept, preserve, and interpret rules with variable constraints is clearly an important aspect of their linguistic competence on langue" (Labov, 1972, p.226). Only the study of language in use will reflect the existence of this ability to operate with variable phenomena.

6. IMPLICATIONS AND CONCLUSIONS

This chapter addresses the contributions that this study makes to the fields of second language acquisition and teaching, its limitations and directions for further studies. These discussions are then followed by a conclusion to the thesis. Let us start by addressing the implications of this study for SLA.

6.1 Implications for Second Language Acquisition

This study investigated the development of sC onset clusters in the English of Hispanophone learners. Although previous studies have identified that preceding phonological environments and the sonority relationship among members of the clusters influence the development of these sequences (Carlisle, 1991a; 1991b; Abrahamsson, 1999; Cutillas-Espinoza, 2002), no study had adopted a variationist perspective to explain the phenomenon. In this context, the present study makes an original contribution to the field of SLA by: (1) adopting a variationist perspective to explain intrinsically variable phenomena, in which both linguistic and extralinguistic factors are used; (2) incorporating three levels of proficiency within a cross-sectional study; (3) incorporating a two-level stylistic distinction, which allows the collection of both controlled and naturalistic data; and (4) using stochastic Optimality Theory as a tool to analyze the variable data encountered in the development of sC clusters.

More importantly, the results obtained in this investigation also contribute to the literature by showing the effects that sonority sequencing has on second language phonological acquisition. Despite some inconsistencies (e.g., the behavior of s+liquid sequences with respect to [e]-epenthesis), the findings presented here seem to confirm the

hypothesis that sonority sequencing is a determining factor in the development of second language syllable structure: in general, structures that abide to sonority are acquired before those that do not.

The following section addresses the pedagogical contributions of the study to the field of second language teaching.

6.2 Pedagogical Implications

The findings obtained in this study can help raise teachers' awareness for the factors involved in the variable acquisition of sC onset clusters by Hispanophone learners of English, and to the extent that the results obtained here are generalizable, to other English learners whose L1s are characterized by the avoidance of these clusters (e.g., Arabic, Bengali, European and Brazilian Portuguese).

Increased teachers' awareness will lead to an appropriate selection and design of materials and tasks that target the more difficult contexts for the surfacing of sC clusters. For instance, teachers will be made aware that not all sC clusters deserve the same amount of attention during pronunciation activities. If the results presented here accurately represent the development of these clusters, s+stops and s+liquids should receive more careful attention in pronunciation activities because they are less likely to be produced accurately. s+nasals, on the other hand, should receive less emphasis because they are more likely to be produced in a target-like manner.

Additionally, teachers will be aware that the production of sC clusters increases with increased exposure to the second language and that, accordingly, "errors" are part of the learning process: change in phonology is ongoing, predictable and systematic, and

determined by a variety of linguistic and extralinguistic factors. For instance, when a beginning student says "I cleaned that [e]stove", the teacher ought to understand that this is simply a phrase in the development of the target language and, more importantly, that the sC cluster involved in the utterance (i.e., s+stop) is of particular difficulty to the language learner, especially because it is preceded by a consonant (e.g., [t] in tha[t] [e]stove). With sufficient training in the L2, the learner will in time master the phonological system that characterizes the target language.

Let us now address the limitations encountered in this study.

6.3 Limitations and further studies

One of the limitations of the present study is that word frequency effects were not taken into consideration in the investigation of English sC onset clusters. Future studies on the development of these clusters might incorporate word frequency in the L2 as an extralinguistic variable, and could make use of audio recordings of the sC cluster learners have been exposed to at a particular language institute to investigate how the L1 transfer patterns (i.e., [e]-epenthesis) are influenced by these factors.

Another limitation of the present study concerns generalizability. Specifically, the unexpected finding that intermediate and advanced learners exhibited monostylism and beginning learners showed stylistic distinctions (albeit small) may just be applicable to EFL classrooms that adopt a communicative approach to language teaching. More precisely, data collected in a communicative classroom may not be generalized to other L2 classrooms because learners' frequent exposure to certain stylistic environments (e.g., informal interviews) could possibly alter the L1 transfer patterns (e.g., [e]-epenthesis).

Despite this limitation, the present study did identify the existence of a relationship between preceding vocalic environments, high proficiency, formal stylistic environments, and the increased production of sC clusters.

A third limitation relates to the number of participants included in this study. According to Seliger and Shohamy (1989), when a study includes a small group of participants, each participant exerts a greater influence on the performance of the group as a whole. The present study initially expected to collect data from 30 EFL learners but difficulties when recruiting participants allowed the collection of data from only 23 learners. This suggests that the inclusion of a larger number of participants might reveal findings different from those obtained in this study.

A fourth limitation concerns the number of tokens collected in the informal task (i.e., 11% of the total collected in the study). Possibly the small percentage of tokens from the informal task influenced the results obtained in the study because it is casual conversation that reveals the learner's true knowledge and ability (Tarone, 1989).

Lastly, the incorporation of two levels of formality in the investigation of sC clusters allowed for the collection of a larger percentage of tokens in the formal task. Perhaps, the adoption of a three-level distinction in a formality hierarchy (i.e., lists of words, lists of phrases, and a picture description task) will allow the collection of a larger amount of tokens in informal environments and might therefore reveal more marked stylistic distinctions across proficiency levels.

6.4 Conclusions

The present study was the first to investigate the variable production of sC clusters in the speech of Hispanophone EFL learners from a variationist perspective. The results revealed that the development of these clusters is systematic and partly comparable to the results obtained in other second language and first language acquisition research. For instance, the statistical results presented here provide further support for the cross-linguistic observations that: (1) preceding consonants and pauses tend to behave in a similar fashion; (2) preceding vocalic environments facilitate the production of sC clusters in the IL grammars of Hispanophone learners of English; and (3) s+stop onset clusters are the most difficult to acquire.

In addition, the results from the extralinguistic factor *proficiency* are consistent with previous sociolinguistic and variationist studies on SLA. For instance, the frequency of L1 transfer features (e.g., [e]-epenthesis) decreased overtime and across proficiency levels, with results that conform to those predicted by the OPM model and to a number of variationist studies on second language acquisition of phonology.

In sum, the results obtained in this investigation demonstrated that the variable acquisition of sC clusters is triggered by a combination of both linguistic (i.e., *preceding phonological environment*, *sC sonority*) and extralinguistic factors (i.e., *proficiency*).

For the analysis of variability in learner speech, this thesis adopted a stochastic version of the framework of Optimality Theory (i.e., the Gradual Learning Algorithm proposed by Boersma et al, 2001). In the analysis, it was proposed that differences in the ranking values assigned to the constraints DEP-IO and *sC across grammars (i.e., across proficiency levels) accounts for the variable production of sC clusters. More importantly,

this stochastic analysis was also able to accurately encode frequency effects into the grammars involved without the need to resort to different grammars. Within stochastic OT, variation within a grammar is a consequence of the distribution of crucially ranked constraints within a strictness scale.

Previous studies on the development of sC clusters suggest that the behavior of these clusters is not uniform. Some studies indicate that sC production is determined by sonority markedness (Carlisle, 1991a, 1991b; Cutillas-Espinoza, 2002); a pattern that is only partly supported in this study due to the behavior of s+liquid clusters, which unexpectedly favored the occurrence of [e]-epenthesis. Another study, however, has found that s+liquids are as difficult to acquire as s+stops (e.g., Abrahamsson, 1999), a finding that is consistent with the results obtained in this investigation. In addition, it has been observed that the speech of Spanish-English bilingual children exhibits a high percentage of [e]-epenthesis application before s+nasals and s+stops (Yavaş, 2005), a pattern that seems to differ from the speech of adult L2 learners, as was shown in this and previous studies. Noticeably, the findings presented and analyzed in this thesis and the previous studies discussed above all share the fact that s+stop onset clusters are the most difficult sequences to acquire, which explains why the Bengali character Chanu is unable to produce the cluster [sk] in "skating" in the following passage from the novel *Brick Lane* by Monica Ali (2003, p.23):

"What is this called?" said Nazneen.

Chanu glanced at the screen. "Ice skating," he said, in English.

"Ice e-skating," said Nazneen.

"Ice skating," said Chanu.

"Ice e-skating."

"No, no. No e. Ice skating. Try it again."

Nazneen hesitated.

"Go on!"

"Ice es-kating," she said, with deliberation.

Chanu smiled. "Don't worry about it. It's a common problem for Bengalis. Two consonants together causes a difficulty. I have conquered this issue after a long time. But you are unlikely to need these words in any case."

"I would like to learn some English," said Nazneen.

(Chanu is a Bangladeshi immigrant who has lived in London for a number of years. Nazneen is his young wife whom Chanu has recently brought back to London from Bangladesh and who doesn't speak English).

References

Abrahamsson, N. (1999). Vowel epenthesis of /sC(C)/ onsets in by Spanish/Swedish interphonology: a longitudinal case study. *Language Learning*, 49 (3), 473-508.

Adamson, H.D. (1988). *Variation theory and second language acquisition.* Washington, D.C.: Georgetown University Press.

Antilla, A. (1997). Deriving variation from grammar: A study of Finnish genitives. In F. Hinskens, R. Van Hout, & L. Wetzels (Eds.). *Variation, Change and Phonological Theory* (pp.35-68). Amsterdam: John Benjamins.

Antilla, A. (2002). Variation and Phonological Theory. In J.K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.). The *handbook of language variation and change* (pp.206-243). Malden, Massachusetts: Blackwell Publishers Ltd.

Archibald, J., & Libben. (1995). *Research Perspectives on Second Language Acquisition.* Missisauga, Ontario: Copp Clark Ltd.

Beebe, L. (1977). The influence of the listener on code-switching. *Language Learning*, 27 (2), 331-339.

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32, 45-86.

Boersma, P., & Weenink (2004). Praat, a system for doing phonetics by computer. Retrieved September 5, 2005 from http://www.praat.org

Bunta, F., & Major, R. (2004). An Optimality Theoretic account of Hungarian ESL learners acquisition of /e /and/æ/. *IRAL: International Review of Applied Linguistics in Language Teaching*, 42 (3), 277-299.

Cardoso, W. (1999). A quantitative analysis of word-final /r/-deletion in Brazilian Portuguese. *Linguistica Atlantica*, 21, 13-52.

Cardoso, W. (2001). Variation patterns in regressive assimilation in Picard. *Language Variation and Change*, 13, 305-341.

Cardoso, W. (2004). The variable acquisition of English word-final stops by Brazilian Portuguese speakers. *Proceedings of the 7th Generative Approaches to Second Language Acquisition Conference (GASLA 2004)*. Somerville, MA: Cascadilla Proceedings Project.

Carlisle, R. (1991a). The influence of syllable structure universals on the variability of interlanguage phonology. In A. Volpe (Ed.). *The Seventeenth LACUS Forum 1990* (pp.135-145). Lake Bluff, IL: Linguistic Association of Canada & the United States.

Carlisle, R. (1991b). The influence of environment on vowel epenthesis in Spanish/ English interphonology. *Applied Linguistics*, 12 (1), 76-95.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (1996). *Teaching Pronunciation: a reference for teachers of English to speakers of other languages*. New York, N.Y.: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cutillas-Espinoza, J.A. (2002). *Sonority and constraint interaction: the acquisition of complex onsets by Spanish learners of English*. Retrieved June 28, 2004 from http://www.uv.es/anglogermanica/2002-1/cutillas.htm#_2_Sonority:_Universal.

Dickerson, L. (1975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly*, 9, 401-407.

Dickerson, L., & Dickerson, W. (1977). Interlanguage phonology: current research and future directions. In S. Corder & E. Roulet (Eds.). *The notions of simplifications, interlanguages and pidgins and their relation to second language learning* (pp.18-29). Paris: AIMAV/Didier.

Eckman, F. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning*, 27, 315-330.

Escartin, C. (to appear). The development of sC onset clusters in Mexican Spanish English. *In Proceedings of the XIX Journées de Linguistique.* Quebec: Laval University.

Fasold, R. (1984). Variation theory and language learning. In P. Trudgill (Ed.). *Applied Sociolinguistics* (pp. 245-261). London: Academic Press.

Gatbonton, E. (1978). Patterned phonetic variability in second language speech: a gradual diffusion model. *Canadian Modern Language Review*, 34, 335-347.

Giegerich, H. (1992). *English phonology.* Cambridge, MA: Cambridge University Press.

Goad, H., & Rose, Y. (2004). Input elaboration, head faithfulness and evidence for representation in the acquisition of left-edge clusters in West Germanic. In R. Kager, J. Pater, & W. Zonneveld (Eds.). *Constraints in phonological acquisition* (pp.109-157). Cambridge, MA: Cambridge University Press.

Harris, J. (1983). *Syllable structure and stress in Spanish: A non-linear analysis.* Cambridge, MA: MIT Press.

Harris, J. (1989). Our present understanding of Spanish syllable structure. In P.C. Bjarkman, & R.M. Hammond (Eds.). *American Spanish pronunciation* (pp.151-169). Washington: D.C: Georgetown University Press.

Harris, J. (1994). *English sound structure*. Cambridge, MA: Blackwell Publishers.

Hancin-Bhatt, B., & Bhatt, R.M. (1997). Optimal L2 syllables: interactions of transfer and developmental effects. *Studies in Second Language Acquisition*, 19, 331-378.

Hayes, B., Tesar, B., & Zuraw, K. (2003). OTsoft 2.1 software package. Retrieved on August 5, 2004 from http://www.linguistics.ucla.edu/people/hayes/otsoft

John, Paul (to appear). Sur la distribution des h épenthétiques dans l'interlangue des apprenants de l'anglais langue seconde. *In Proceedings of the XIX Journées de Linguistique*. Quebec: Laval University.

Kiparsky, P. (1993). Variable rules. *Paper presented at the Rutgers Optimality Workshop*. New Brunswick, NJ.

Krashen, S. (1981). *Second language acquisition and learning*. Oxford: Pergamon Press.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language*, 45, 715-762.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lin, Y. (2003). Interphonology variability: sociolinguistic factors affecting L2 simplification strategies. *Applied Linguistics*, 24 (4), 439-464.

Major, R. (2001). *Foreign Accent. The Ontogeny and Phylogeny of Second Language Phonology*. New Jersey: Lawrence Erlbaum Associates

Major, R. (2004). Gender and stylistic variation in second language phonology. *Language Variation and Change*, 16, 169-188.

McCarthy, J., & Prince, A. (1995). Faithfulness and reduplicative identity. In J. Beckman, L. Walsh-Dickey and S. Urbanczyk (Eds.). University of Massachusetss

Occasional Papers in Linguistics 18: Papers in Optimality Theory (pp. 249-384). Amherst, MA: GLSA.

Preston, D. (1996). Variationist perspectives on second language acquisition. In R. Bayley & D. Preston (Eds.). *Second Language Acquisition and Linguistic Variation* (pp.1-45). Amsterdam: Benjamins.

Prince, A., & Smolensky, P. (1993). *Optimality: Constraint interaction in generative grammar*. Cambridge, Massachusetts: MIT Press.

Reynolds, W. (1994). *Variation and phonological theory*. PhD thesis, University of Pennsylvania.

Robinson, J., Lawrence, H., Tagliamonte, S. (2001). *Goldvarb 2001: A multivariate analysis application for windows – User's manual*. Retrieved August 5, 2004 from http://www.york.ac.uk/depts/lang/webstuff/goldvarb.

Seliger, H.W., & Shohamy, E. (1989). *Second Language Research Methods*. Oxford: Oxford University Press.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10, 209-231.

Selkirk, E. (1972). *The phrase phonology of English and French*. PhD thesis, MIT.

Skehan, P. (1989). *Individual differences in second-language learning*. London: Edward Arnold.

Spencer, A. (1996). *Phonology*. Cambridge, MA: Blackwell Publishers Inc.

Tarone, E. (1985). Variability in Interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning*, 35, 373-403.

Tarone, E. (1989). Accounting for style-shifting in interlanguage. In S. Gass, C. Madden, D. Preston, & L. Selinker (eds.). *Variation in Second Language Acquisition: Psycholinguistic Issues* (pp.13-21). Clevendon, UK: Multilingual Matters.

Tesar, B., & Smolensky, P. (1995). The learnability of Optimality Theory. In R. Aranovich, W.Byrne, S. Preuss, and M. Senturia (eds.). *Proceedings of the 13th West Coast Conference on Formal Linguistics* (pp.122-137). Stanford, CA: CSLI Publications.

Tesar, B., & Smolensky, P. (1998). Learnability of Optimality theory. *Linguistic Inquiry*, 29, 229-268.

Winford, D. (1992). Back to the past: the BEV/Creole connection revisited. *Language Variation and Change*, 4 (3), 311-357.

Yavaş, M., & Someillan, M. (2005). Patterns of acquisition of /s/-clusters in Spanish-English bilinguals. *Journal of Multilingual Communication Disorders*, 3(1), 50-55.

Young, R. (1991). *Variation in Interlanguage Morphology*. New York, N.Y.: P. Lang Publishing Inc.

Young, R., & Bayley R. (1996). VARBRUL analysis for second language acquisition research. In R. Bayley and D. Preston (eds.), *Second Language Acquisition and Linguistic Variation*. Amsterdam: John Benjamins (253-306).

Appendix A. Questionnaire

Name:_____

Instructions: Answer the following questions by marking an (x) in the option that best describes you.

1. What is your age group?

- ☐ 21-25
- ☐ 26-30

2. Gender

- ☐ female
- ☐ male

3. What is your English level at this school?

- ☐ Beginner (1-2)
- ☐ Intermediate (3-4)
- ☐ Advanced (5-6)

4. From the statements provided below select the one that best describes you.

- ☐ I really enjoy learning English
- ☐ I somewhat like learning English
- ☐ I do not like learning English at all

5. How long have you studied English?

- ☐ 0-1 year
- ☐ 1-2 years
- ☐ 2-3 years
- ☐ 3 or more years

6. Choose the option that best describes your exposure to English outside the classroom.

- ☐ In an English speaking country
- ☐ With a private teacher
- ☐ With English teachers
- ☐ Other:
  Specify: _____
- ☐ None of the above

7. Choose the statement that best describes the time you have devoted to learning English pronunciation.

- ☐ I have never studied English pronunciation
- ☐ I studied English pronunciation for less than six months
- ☐ I studied English pronunciation for more than six months

8. If you answered positively to the previous question, where did you learn about English pronunciation?

- ☐ In the English classroom
- ☐ In a special course
- ☐ With a private teacher
- ☐ On my own (for example: on my computer, surfing the internet, with a CD or tape, with special software)

Thank you for you participation!

Appendix B. Grammaticality judgment task

Name: _____

Instructions: Read aloud the sentence that you consider as grammatically correct.

1. (a) That big room smells like books
   (b) That big room smells as books

2. (a) A lot of fire came out of the small stove
   (b) A lot of fire camed out of the small stove

3. (a) Paul soon slapped the violent man, who stopped the school play
   (b) Paul was soon slapped the violent man, who stopped the school play

4. (a) Speeding can cause a lot of accidents and deaths
   (b) Speeding can be cause accidents and deaths

5. (a) John and Mary like to snuggle when they dance slowly
   (b) John and Mary like to snuggling when they dance slowly

6. (a) In spite of the terrible snow storm, Kim and Steve played in the backyard
   (b) Because of the terrible snow storm, Kim and Steve played in the backyard

7. (a) Two massive spaceships were constructed by NASA
   (b) Two massive spaceships was constructed by NASA

8. (a) "study hard" said the teacher to her smart pupils
   (b) "study hard" sayed the teacher to her smart pupils

9. (a) George Bush hit a paparazzi who took forbidden snapshots of his daughter
   (b) George Bush hit a paparazzi who taked forbidden snapshots of his daughter

10. (a) Stacey looked spectacular with her pink skirt
    (b) Stacey look spectacular with her pink skirt

11. (a) My brother loves guava smoothies
    (b) My brother loving guava smoothies

12. (a) John will soon stay home and sleep late
    (b) John is going soon stay home and sleep late

13. (a) Greet costumers with a big smile says my boss
    (b) Greeting costumers with a big smile says my boss

14. (a) The house stinks because of the huge skunk that came in yesterday
    (b) The house stinks because the huge skunk that came in yesterday

15. (a) More spiders keep sneaking in my bedroom
    (b) More spiders keep sneaking into my bedroom

16. (a) More slim girls entered the beauty contest this year
    (b) More slim girls entered the beautiful contest this year

17. (a) Smashing pumpkins made a fun sketch five years ago
    (b) Smashing pumpkins made a fun sketch five years before

18. (a) Pa scolded me because my friends and I are skipping classes
    (b) Pa scolded me why my friends and I are skipping classes

19. (a) Less smog is what polluted cities need
    (b) Less smog is what pollution cities need

20. (a) Five snakes were slithering around the spare room in my house
    (b) Five snakes were slither around the spare room in my house

21. (a) Smiling the four-year old took the toy
    (b) Smiling the four-year old tooked the toy

22. (a) Stop offering free slices of pie
    (b) Stop offers free slices of pie

23. (a) Spencer won free skating lesson at a radio contest
    (b) Spencer wined free skating lesson at a radio contest

24. (a) Scan the reading said the teacher in blue spandex pants
    (b) Scan the reading say the teacher in blue spandex pants

25. (a) The law states that it is illegal to rob banks
    (b) The law states that it is illegal rob banks

26. (a) Slippery floors are very dangerous
    (b) Slippery floors are very danger

27. (a) My toe slowly began swelling after I hit myself
    (b) My toe slowly began to swell after I hit myself

28. (a) Snow, skate are two verbs that always go together in winter
    (b) Snow, skate are two verbs that always goes together in winter

29. (a) I like to draw snails, smurfs, and a paw slightly blue
    (b) I likes to draw snails, smurfs, and a paw slightly blue

30. (a) Slam, spiky hair, and gray lipgloss were a big trend in the 90s
    (b) Slam, spiky hair, and gray lipgloss have been a big trend in the 90s

31. (a) Free snacks are being offered at the meeting
    (b) Free snacks is being offered at the meeting

32. (a) Draw spiders inside the box, ordered the teacher
    (b) Draw spider inside the box, ordered the teacher

33. (a) Slice the turkey carefully, then serve it with some potato salad, and enjoy
    (b) Slice the turkey careful, then served with some potato salad, and enjoy

34. (a) Skin care is essential to avoid aging
    (b) Skin cared is essential to avoid aging

35. (a) Snobby people are not always nice to other people
    (b) Snob people are not always nice to other people

36. (a) Sneezing is the first symptom of a cold
    (b) Sneezing is the first symptomatic of a cold

Appendix C. sC clusters and preceding environments

Tables 1, 2, and 3 contain the distribution of preceding environments (i.e., consonants, vowels, and pauses) across the 36 pairs of sentences included in the formal task. Table 4 shows the onset clusters (i.e., s+liquid, s+nasal, and s+stop) included in the task.

Table 1. Onset clusters and preceding consonantal environments

Cluster	Preceding consonant	Sentence #	Total
/st/	/l/	2	4
	/d/	6	
	/n/	12	
	/z/	14	
/sm/	/m/	1	4
	/r/	8	
	/g/	13	
	/s/	19	
/sp/	/n/	6	4
	/v/	7	
	/t/	10	
	/r/	15	
/sl/	/n/	3	4
	/z/	5	
	/d/	12	
	/r/	20	
/sn/	/l/	6	4
	/n/	9	
	/p/	15	
	/v/	20	
/sk/	/k/	10	4
	/dʒ/	14	
	/n/	17	
	/r/	18	
			24

Table 2. Onset clusters and preceding vocalic environments

Cluster	Preceding vowel	Sentence #	Total
/st/	/ʊ/	3	3
	/ow/	6	
	/ɔ/	25	
/sm/	/ə/	2	3
	/iy/	7	
	/ɑ/	11	
/sp/	/ə/	2	3
	/uw/	24	
	/ɔ/	32	
/sl/	/iy/	22	3
	/ow/	27	
	/ɔ/	29	
/sn/	/ə/	5	3
	/ɔ/	29	
	/iy/	31	
/sk/	/ə/	3	3
	/ɑ/	18	
	/iy/	23	
			18

Table 3. Onset clusters and preceding environments (pauses)

Cluster	Sentence #	Total
/st/	8 10 22	3
/sm/	17 21 29	3
/sp/	4 23 30	3
/sl/	26 30 33	3
/sn/	28 35 36	3
/sk/	24 28 34	3
		18

Table 4. Words with onset clusters

Cluster	Words	Total
/st/	stove, stopped, storm, Steve, Stacey, stay, stinks, stop, states, study	10
/sm/	smell, small, smart, smoothies, smile, smashing, smog, smiling, smart, smurfs	10
/sp/	speeding, spite, spaceships, spectacular, spiders, spare, Spencer, spandex, spiky, spiders	10
/sl/	slapped, slowly, sleep, slither, sleigh, slowly, slightly, slam, slice, slice	10
/sn/	snuggle, snow, snapshots, sneaking, snakes, snow, snails, snacks, snobby, sneezing	10
/sk/	school, skirt, skunk, sketch, scolded, skipping, skating, scan, skate, skin	10
		60

Appendix D. Sample Questions: Informal Interview

01. What is your name?

02. How old are you?

03. Do you have any brothers or sisters? How old are they?

04. Where were you born? (When did you move to Aguascalientes? Why did you come here?)

05. Why are you studying English? Do you enjoy studying English?

06. For how long have you studied English?

07. Do you like Aguascalientes? Why/why not?

08. Where are your parents from? (When did they move here? Why did they come here?)

09. What do you do when you aren't at Interlingua?

    - Do you work? (What kind of work do you do?)

    - Are you looking for work? (What kind of work? What kind of work have you done in the past?)

10. What are you studying? What kind of work would you like to do in the future?)

11. What do you like to do in your free time?

    - Do you have any special interests or hobbies?

    - Do you enjoy listening to music or watching films or playing sports or reading or surfing the internet or clothes shopping? What is your favourite film/tv program/book/cd?

    - What do you do with your friends?

12. What did you do last weekend? What are you going to do next Christmas?

13. What do you think your life will be like in ten years time?

    - Will you be married with children? (How many children do you want to have?)

- Will you be working? (What will your job be?)
- Will you be living in another place, another city or country?

14. Is there another country you would like to live in or visit? (Why do you want to go there?)

15. Of all the places you have visited, which place do you like the most?

16. Is there a person you really admire? (It can be a famous person or not. Is there someone you would really like to meet?)

17. Tell me something you can do in English now that you couldn't do before you started studying here.

18. Tell me something you can't do in English now but that you want to do in English in the future.

19. Tell me about the last movie you saw. How was it? What was the movie about?

20. Is there anything else you would like to tell me about yourself? Something I haven't asked about?

21. Do you have any questions for me? Is there anything you would like to know about myself?

Appendix E. Total number of tokens per participant[11]

Participant	[e]	sC	Total	%
1	36	30	66	4
2	5	66	71	4
3	28	35	63	4
4	48	22	70	4
5	7	64	71	4
6	11	61	72	4
7	29	39	68	4
8	51	13	64	4
9	20	46	66	4
0	2	62	64	4
!	10	56	66	4
#	10	55	65	4
$	59	5	64	4
%	46	21	67	4
^	23	45	68	4
&	2	62	64	4
*	21	43	64	4
-	4	65	69	4
=	45	19	64	4
_	32	34	66	4
+	6	57	63	4

[11] A total of 60 sC tokens per participant were elicited through the grammaticality judgment task.

Appendix F. Total number of tokens per proficiency level

Proficiency	[e]	sC	Total	%
a	96	386	482	31
i	163	367	530	34
b	326	189	515	33
Total	585	942	1527	

Appendix G. Coding scheme for Goldvarb analysis

Dependent Variable		0	sC clusters
		1	[e]-epenthesis
Independent variables	(1) sC sonority	L	s + liquid
		N	s + nasal
		S	s + stop
	(2) Preceding environment	l	liquid
		n	nasal
		s	stop
		f	fricative
		p	pause
		v	vowel
	(3) Proficiency	b	beginner
		i	intermediate
		a	advanced
	(4) Level of formality	F	formal
		I	informal
	(5) Participants	1, 2, 3, 4 ..	

Appendix H. Final Goldvarb results for the relevant groups*

Factor groups	Factors	/e/ epenthesis			
		N	%	p	input
(1) sC sonority	s+liquid	101	40	.52	.34
	s+nasal	153	33	.42	.25
	s+stop	331	40	.54	.36
(2) Preceding environment	Consonant	274	43	.59	.40
	Vowel	136	28	.34	.20
	Pause	175	41	.55	.37
(3) Proficiency level	Advanced	96	19	.32	.18
	Intermediate	163	30	.44	.27
	Beginner	326	63	.71	.54
(4) Formality level	Formal	522	38	.49	.31
	Informal	63	40	.58	.40

* The probability weights (rounded up to two digits) result from the final Goldvarb analysis (Binomial 1-level) conducted without the speaker whose grammar was near categorical, with the removal of the redundant factor group *participants*, and with the recoding of the preceding consonants (i.e., liquids, nasals, stops, fricatives) as one single-factor 'consonants'.