INTERACTION, CORRECTIVE FEEDBACK, AND THE DEVELOPMENT OF LEXICAL STRESS

Özgür Parlak, M.A.

The Department of Linguistics and English Language

Lancaster University

Thesis submitted for the degree of Doctor of Philosophy in Linguistics

October, 2018

Statement of Ownership

I, Özgür Parlak, confirm that the research presented in this thesis is my own and it has not been submitted for the award of a higher degree elsewhere. I also confirm that information obtained from other sources has been acknowledged.

Özgür Parlak

INTERACTION, CORRECTIVE FEEDBACK, AND THE DEVELOPMENT OF LEXICAL STRESS

Özgür Parlak, M.A.

October, 2018

ABSTRACT

Recasts are probably the most commonly studied type of correct feedback by second language acquisition (SLA) researchers and there is substantial evidence that they facilitate different types of L2 development (Li, 2010; Mackey & Goo, 2007; Sheen & Ellis, 2011). However, to date, their impact on learners' phonological development has received relatively little attention. In order to fill this gap, the current study examined the effects of recasts on the development of lexical stress in L2 English. Following a pretest-posttest design, 68 L1 Arabic speakers were randomly assigned to intervention and control groups. The pre- and posttests comprised sentence-reading and decision-making tasks that contained three-syllable words with stress on the second syllable as the target vocabulary. All learners participated in an interactive role-play task, with learners in the intervention group receiving a recast following misplaced primary stress during the role-play task and learners in the control group not receiving any form of corrective feedback.

Acoustic analyses of learners' primary stress placement focused on syllable duration, pitch, and intensity because of their established role as the main correlates of stress in English (Beckman, 1986; Cutler, 2008; Lieberman, 1960). Results demonstrated that the intervention group's realizations of second syllables at the posttest exhibited statistically longer duration and higher pitch than their pretest

productions, whereas the control group did not show any gains. Furthermore, acoustic analyses of the target words produced by the intervention group showed that the gains were limited to the words that received a recast. In sum, the findings showed a positive effect for recasts on the development of primary stress using acoustic analytical tools. Thus, this thesis contributes to the growing body of SLA research by revealing that recasts can promote development of primary stress placement, hence L2 phonology, a relatively understudied area within the interactionist strand of SLA.

ACKNOWLEDGEMENTS

I consider my PhD journey to be the outcome of a concerted effort by several people. First and foremost, I am truly grateful to my dissertation committee, Sam Kirkham, Alison Mackey, and Jenefer Philp, for their guidance and encouragement during every step of this research project. I have been blessed with three enthusiastic mentors who have always been there for me when I need their advice and support. It goes without saying that the writing of this dissertation would not have been possible without their inspiring mentorship and belief in me.

I would like to thank my dear friend Dr. Nicole Ziegler of the University of Hawaiʻi at Mānoa for our discussions and her invaluable input during the piloting stage of this dissertation. I would also like to extend my heartfelt thanks to Dr. Kim McDonough of Concordia University for instilling the love of SLA in me; and Dr. Doug Biber, Dr. Bill Grabe, Dr. Okim Kang, Dr. Randi Reppen, and Dr. Fredricka Stoller of Northern Arizona University for helping me expand my knowledge in Applied Linguistics and encouraging me to pursue a doctorate.

I am thankful to the American University of Sharjah for providing me with a stimulating academic environment, giving me the time I needed to work on this dissertation, and funding my overseas conference presentations that discussed different aspects of this research project. The support I have received from my department and college has been instrumental in completion of this dissertation.

Working on a PhD can be isolating at times with long periods of seclusion. I am obliged to my wife, Salma, for her patience and understanding during my periods of isolation which allowed me to channel my energy and time toward this project. Last but not least, I would like to thank my parents, Ali and Akgül, for everything they have done for me throughout the years; and I would like to dedicate this dissertation to my mother, for whom my education and well-being have always been a top priority.

TABLE OF CONTENTS

ABSTRACT 1
ACKNOWLEDGEMENTS 3
LIST OF TABLES 6
LIST OF FIGURES 8
CHAPTER 1: INTRODUCTION 10
CHAPTER 2: THE INTERACTION APPROACH 17
2.1. Input 18
2.2. Output 21
2.3. Noticing and attention 23
2.4. Salience 26
2.5. Corrective feedback 30
2.6. Recasts 36
CHAPTER 3: PHONOLOGY AND SLA 43
3.1. Theory driven L2 phonology studies 47
3.2. Pronunciation instruction studies 52
3.3. Interactionist L2 phonology studies 58
3.4. The case for recasts and lexical stress 63
CHAPTER 4: LEXICAL STRESS 70
4.1. Segmentals and suprasegmentals 70
4.2. Accent and stress 70
4.3. Lexical stress in English and Arabic 77
4.4. Importance of lexical stress for L2 learners 82
CHAPTER 5: THE PILOT STUDY 90
5.1. Method 90
5.1.1. Research questions 90
5.1.2. Participants 90
5.1.3. Language tasks 91
5.1.4. Target words 96
5.1.5. Procedure 97
5.2. Acoustic analysis 99
5.2.1. Coding 99
5.2.3. Mixed-effects models 100
5.2. Results of acoustic analyses 100
5.3. Auditory analyses 101
5.4. Discussion and changes for the main study 102
CHAPTER 6: METHODS AND STUDY DESIGN 105
6.1. Research questions 105
6.2. The context 105
6.3. Participants 106
6.3.1. Learners 107

6.3.2. Listeners 110
6.4. Instruments 112
6.4.1. Language tasks 112
6.4.2. Target words 113
6.4.3. Exit survey 117
6.4.4. Listener judgment experiment 117
6.5. Study design 118
6.6. Procedure 123
6.6.1. Data collection from learners 123
6.6.2. Data collection from listeners 126
6.7. Acoustic analysis 129
6.7.1. Coding of acoustic measures 129
6.7.2. Statistical analysis of acoustic measures 138
6.7.3. Model fitting for acoustic analysis 1 144
6.7.4. Model fitting for acoustic analysis 2 148
6.7.5. Calculation of goodness-of-fit measures 151
6.7.6. Pairwise comparisons 155
6.8. Analysis of listener judgments 156
CHAPTER 7: RESULTS 157
7.1. Acoustic analysis 1 157
7.1.1. Duration 157
7.1.2. Intensity 167
7.1.3. Pitch 175
7.2. Acoustic analysis 2 183
7.2.1. Duration 183
7.2.2. Intensity 187
7.2.3. Pitch 190
7.3. Summary of acoustic analyses 193
7.4. Listener judgments 203
7.5. Exit survey 204
CHAPTER 8: DISCUSSION AND CONCLUSION 209
8.1. Summary of answers to research questions 210
8.2. Discussion of findings 212
8.2.1. Changes in acoustic correlates 212
8.2.2. The benefits of recasts for the development of lexical stress 216
8.2.3. Listener judgments 220
8.2.4. Pedagogical implications 226
8.2.5. Limitations and future directions 231
REFERENCES 243
APPENDICES 293

LIST OF TABLES

Table 2.1 Single-move recasts 39
Table 2.2 Multi-move recasts 40
Table 5.1 Target words used in the pilot study 97
Table 6.1 Learner demographics by group 110
Table 6.2 Listener demographics 112
Table 6.3 Target words 116
Table 6.4 Syllable boundaries of the target words 130
Table 6.5 Comparison of R^2 values obtained for Acoustic Analysis 1 154
Table 6.6 Comparison of R^2 values obtained for Acoustic Analysis 2 154
Table 7.1 Descriptive statistics for proportional difference in duration by experimental condition 157
Table 7.2 Linear mixed-effects model estimates for proportional difference in duration by experimental condition 158
Table 7.3 Predicted pretest and posttest means for proportional difference in duration by experimental condition 159
Table 7.4 Pairwise pretest-posttest contrasts for proportional difference in duration by experimental condition 159
Table 7.5 Descriptive statistics for proportional difference in intensity by experimental condition 167
Table 7.6 Linear mixed-effects model estimates for proportional difference in intensity by experimental condition 168
Table 7.7 Predicted pretest and posttest means for proportional difference in intensity by experimental condition 168
Table 7.8 Pairwise pretest-posttest contrasts for the proportional difference in intensity by experimental condition 169
Table 7.9 Descriptive statistics for proportional difference in pitch by experimental condition 175
Table 7.10 Linear mixed-effects model estimates for proportional difference in pitch by experimental condition 176
Table 7.11 Predicted pretest and posttest means for proportional difference in pitch by experimental condition 176
Table 7.12 Pairwise pretest-posttest contrasts for proportional difference in pitch by experimental condition 177
Table 7.13 Descriptive statistics for proportional difference in duration by feedback provision 184
Table 7.14 Linear mixed-effects model estimates for proportional difference in duration by feedback provision 184
Table 7.15 Predicted pretest and posttest means for proportional difference in duration by feedback provision 185

Table 7.16 Pairwise pretest-posttest contrasts for proportional difference in duration by feedback provision 185
Table 7.17 Descriptive statistics for proportional difference in intensity by feedback provision 187
Table 7.18 Linear mixed-effects model estimates for proportional difference in intensity by feedback provision 188
Table 7.19 Predicted pretest and posttest means for proportional difference in intensity by feedback provision 188
Table 7.20 Pairwise pretest-posttest contrasts for proportional difference in intensity by feedback provision 188
Table 7.21 Descriptive statistics for proportional difference in pitch by feedback provision 191
Table 7.22 Linear mixed-effects model estimates for proportional difference in pitch by feedback provision 191
Table 7.23 Predicted pretest and posttest means for proportional difference in pitch by feedback provision 191
Table 7.24 Pairwise pretest-posttest contrasts for proportional difference in pitch by feedback provision 192
Table 7.25 Comparison of gains by group and by feedback provision 199
Table 7.26 Cohen's Kappa analysis of stress placement judgments 203
Table 8.1 Summary of answers to research questions 211

LIST OF FIGURES

Figure 2.1 Model of interaction and learning 31
Figure 6.1 Dialects of Arabic spoken by learners 108
Figure 6.2 Study design 119
Figure 6.3 The procedure for data collection from learners 124
Figure 6.4 The procedure for data collection from listeners 128
Figure 6.5 Marking of syllable boundaries 133
Figure 7.1 95% confidence intervals for the proportional difference in duration by experimental condition 159
Figure 7.2 Pretest (left) and posttest (right) distributions of duration measures by learners in the intervention group 161
Figure 7.3 Pretest (left) and posttest (right) distributions of duration measures by learners in the control group 162
Figure 7.4 Pretest (left) and posttest (right) distributions of duration measures at the word level by learners in the intervention group 164
Figure 7.5 Pretest (left) and posttest (right) distributions of duration measures at the word level by learners in the control group 165
Figure 7.6 Comparison between NS and NNS productions of proportional difference in duration by experimental condition 166
Figure 7.7 95% confidence intervals for proportional intensity difference by experimental condition 169
Figure 7.8 Pretest (left) and posttest (right) distributions of intensity measures by learners in the intervention group 170
Figure 7.9 Pretest (left) and posttest (right) distributions of intensity measures by learners in the control group 171
Figure 7.10 Pretest (left) and posttest (right) distributions of intensity measures at the word level by learners in the intervention group 172
Figure 7.11 Pretest (left) and posttest (right) distributions of intensity measures at the word level by learners in the control group 173
Figure 7.12 Comparison between NS and NNS productions of proportional difference in intensity by experimental condition 174
Figure 7.13 95% confidence intervals for proportional difference in pitch by experimental condition 177
Figure 7.14 Pretest (left) and posttest (right) distributions of pitch measures by learners in the intervention group 178
Figure 7.15 Pretest (left) and posttest (right) distributions of pitch measures by learners in the control group 179
Figure 7.16 Pretest (left) and posttest (right) distributions of pitch measures at the word level by learners in the intervention group 180

Figure 7.17 Pretest (left) and posttest (right) distributions of pitch measures at the word level by learners in the control group 181
Figure 7.18 Comparison between NS and NNS productions of proportional difference in pitch by experimental condition 182
Figure 7.19 95% confidence intervals for proportional difference in duration by feedback provision 185
Figure 7.20 Comparison between NS and NNS productions of proportional difference in duration by feedback provision 186
Figure 7.21 95% confidence intervals proportional intensity difference by feedback provision 189
Figure 7.22 Comparison between NS and NNS productions of proportional difference in intensity by feedback provision 190
Figure 7.23 95% confidence intervals for proportional difference in pitch by feedback provision 192
Figure 7.24 Comparison between NS and NNS productions of proportional difference in pitch by feedback provision 193
Figure 7.25 Correlation between word frequency and gains in duration 197
Figure 7.26 Confidence intervals for Acoustic Analysis 1 199
Figure 7.27 Confidence intervals for Acoustic Analysis 2 200
Figure 7.28 Learners' responses to the item "During the interview task, I focused on pronunciation" by group 204
Figure 7.29 Learners' responses to the item "The interviewee gave me feedback when I made pronunciation mistakes" by group 205
Figure 7.30 Learners' responses to the item "Did the interviewee correct your English during the interview task?" by group 206

CHAPTER 1: INTRODUCTION

Second Language Acquisition (SLA) is a dynamic scientific discipline that investigates questions related to the human capacity to learn a language in addition to one's native language. The SLA field has its roots in the scholarly work that was published around 1950s, which aimed to form a new theory of language learning that was particularly inspired by structural linguistics and behaviorist psychology. Drawing on Skinner's (1957) seminal work on operant conditioning and his view that language learning is a form of habit formation, Lado (1957) introduced contrastive analysis, a method for identifying linguistic difficulties that learners face when learning a second/foreign (L2) language based on the premise that the difficulties faced in L2 are the result of learners' first language (L1) habits. As a result of the interest in contrastive analysis at the time, research effort was directed toward producing structural comparisons between various language pairs with the intention of mapping out which L1 features, or habits, could hinder acquisition in L2; and in the same way, which similarities between the L1 and the L2 could facilitate L2 acquisition. Contrastive analysis, or behaviorist and structuralist approaches in general, did not consider learners themselves as a factor that impacts the language learning process. However, SLA approaches that considered learners as an empty slate did not last long. Corder (1967) was one of the earliest L2 researchers to suggest that L2 learners may have a build-in syllabus similar to the one that children have during the early stages of L1 acquisition and argued that "the key concept in both cases is that the learner is using a definite system of language at every point in his development" (p. 166). These ideas brought about a fundamental shift in focus which moved the discussion within the SLA field from a structural comparison of language pairs to an investigation of learner errors as a window to learners' developmental stages as well as cognitive learning mechanisms. The shift from behaviorism to cognitive approaches developed further by Selinker's (1972) proposal of interlanguage, "a separate linguistic system based on the observable output which results from a learner's attempted production of a TL norm" (p. 214). Over the past four and a half decades, the SLA field has seen a number of major developments and SLA researchers today employ cognitive approaches to a great extent (for a detailed review, see Larsen-Freeman, 2007). The past couple of decades have also been productive in terms of advancements in research methodology and building clearly defined constructs. Methodological advancements have allowed researchers to investigate research questions pertaining to various aspects of L2 learning. So far, SLA researchers have explored issues related to L2 learners' comprehension, interlanguage development, and production from various different perspectives including linguistic (e.g., Gregg, 1996; L. White, 1989), sociocultural (e.g., Lantolf, 2000; Swain & Lapkin, 2002) and cognitive (e.g., Bates & MacWhinney, 1981; DeKeyser, 2001; Gass, 1997; Long, 1983; Pica, 1994) approaches. The substantial amount of research output and key findings produced by countless studies have helped SLA become an established scientific field that contributes to the body of knowledge that guides our understanding of L2 development in naturalistic as well as instructional environments.

Among the various strands of SLA, a prominent one is the interactionist approach. The interactionist approach has its roots in Krashen's (1977, 1982) work on the monitor model and comprehensible input hypothesis as well as the interaction hypothesis that dates back to the works of Hatch (1978) and Long (1981, 1983), which highlighted the need for studying potential facilitative effects of linguistic input and communicative interaction on language development. Their work expanded on

Krashen's (1977, 1982, 1985) comprehensible input hypothesis by acknowledging the role of input but at the same time suggesting that it would not be enough on its own for successful L2 development. The ideas presented by Hatch and Long resonated well with other SLA researchers who contributed to this new research agenda by investigating various aspects of the interaction hypothesis resulting in findings that underscored the role of pushed output (Swain 1985, 1995), negotiation for meaning (Pica, 1996a, 1996b; Varonis & Gass, 1985), and attention that triggers noticing the gap between the target language form and learners' erroneous production (Gass, 1997; Schmidt, 1990; Schmidt & Frota, 1986). In an effort to account for new evidence and findings that shed light on different aspects of the relationship between interaction and language learning, Long (1996) revised his initial proposal and redefined the interaction hypothesis as "negotiation for meaning, and especially negotiation work that triggers interactional adjustments by the native speaker or more competent interlocutor, facilitates acquisition because it connects input, internal learner capacities, particularly selective attention, and output in productive ways" (pp. 451-452). The revised interaction hypothesis stood on four key elements that take place during meaning-oriented communication: input, interaction, feedback, and output. These seminal ideas gave birth to a brand-new strand within the SLA field and have shaped the interactionist research agenda over the past two decades. Currently, it is generally accepted by SLA researchers that the type interaction described by Long is a fundamental component of L2 acquisition (Ellis, 2008; Gass & Mackey, 2006, 2015; Kim, 2017; Spada & Lightbown, 2009). This view is also supported by a number of meta-analyses that have shown that interaction that provides opportunities for negotiation for meaning and provision of corrective feedback facilitates L2 acquisition and that the effect on learning is durable (Keck, Iberri-Shea, Tracy-Ventura, & Wa-

Mbaleka, 2006; Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007; Russell & Spada, 2006). As a result of the significant amount of empirical evidence that had been produced by early 2000s, the interaction hypothesis was no longer just a hypothesis but a complex web of constructs that aims to explain how humans, children or adult, learn a second language. SLA researchers have adopted different terminology when referring to this web of constructs, including *theory* (Carroll & Swain, 1993) and *framework* (Mackey, Abbuhl, & Gass, 2011). More recently, Gass and Mackey (2015) explained that the interaction hypothesis is not a complete theory of SLA but at the same time, in light of the accumulating research findings that support the usefulness of interaction, it is not a hypothesis either. Thus, they suggested using *the interaction approach* when referring to the body of research that investigates the intricacies of the relationship between communicative interaction and L2 acquisition. Following their lead, the term interaction approach will be adopted throughout this dissertation.

Interactionist researchers have stressed that the benefits of interaction are usually subject to the nature and complexity of the target language feature as well as the nature, salience, or the content of feedback received by the learner (Long, 2007; Mackey, Gass, & McDonough, 2000; Pica, 1994). Research to date has shown positive effects of interaction on L2 acquisition by different age groups (Mackey & Sachs, 2012; Mackey & Silver, 2005), in both classroom and laboratory contexts (Gass, Mackey, & Ross-Feldman, 2005), and for a broad range of target language features such as English possessive determiners (Ammar & Spada, 2006), English dative alteration (Carroll & Swain, 1993), Spanish vocabulary (de la Fuente, 2002; Gass & Alvarez-Torres, 2005), simple past tense in English (Ellis, Loewen, & Erlam, 2006; McDonough, 2007), counterfactual past construction in English (Révész, Sachs,

& Hama, 2014), past tense in French (Ayoun, 2001), Spanish gender agreement (Gass & Alvarez-Torres, 2005; Leeman, 2003), question formation in English (Mackey & Philp, 1998; McDonough & Mackey, 2006, 2008), that-trace filter in English (Goo, 2012), Japanese aspectual form -te i-(ru) (Ishida, 2004), Spanish adverb placement (Long, Inagaki, & Ortega, 1998), Japanese locative-initial construction (Iwashita, 2003), English articles (Sheen, 2007, 2008), and French grammatical gender (Lyster, 2004). It is quite clear from these examples, which by no means represent a complete list, that the efficacy of interaction for L2 learning has been put to test through addressing a wide range of research questions and target language features. That said, a substantial amount of interactionist SLA studies conducted so far have particularly focused on the acquisition of morphology and lexis. In comparison, the number of studies that investigated the relationship between interaction and the acquisition of L2 phonology has remained relatively low. Among those few interactionist studies, a study by Trofimovich, McDonough, and Foote (2014) investigated the relationship between interactive alignment facilitated through collaborative tasks and correct placement of lexical stress. Their participants, who came from a wide range of L1 backgrounds, produced a higher number of target-like stress patterns following their interlocutor's target-like lexical stress production. Saito and Wu (2014) investigated the effects of form-focused instruction and corrective feedback on L1 Cantonese speakers' perception of Mandarin tones. They reported that corrective feedback helped participants improve their perception performance only under trained lexical conditions, which they interpreted as a lack of generalizability. In another study that focused on perception, Lee and Lyster (2016a) investigated the effects of form-focused instruction and corrective feedback on L1 Korean speakers' perceptual accuracy of the English /i/-/ɪ/ phonemic contrast. The findings from their study showed that corrective feedback facilitated improved perception on both immediate and delayed posttests. Although these studies adopted an interactionist approach when exploring development of L2 phonology, they did not examine the effects of corrective feedback on L2 speech production. In fact, there are only a handful of studies that specifically focused on L2 phonological development using an interactionist approach. In an attempt to explore this understudied area, Saito and Lyster (2012a) investigated the effects of form-focused instruction and corrective feedback on the development of the approximant /ɹ/ in L2 English. The group of L1 Japanese speakers that received form-focused instruction and corrective feedback showed significant acoustic gains at the posttest in their production of /ɹ/, which was measured against a native-speaker baseline. On the other hand, the group that received form-focused instruction without corrective feedback and the control group did not show any improvement. In another paper that reported the results on the development of /æ/, Saito and Lyster (2012b) observed that instruction coupled with pronunciation-focused recasts led to more native-like production of the high-front vowel /æ/. In a subsequent study, Saito (2015a) found that form-focused instruction alone was able to promote L1 Japanese learners' perception and production of /ɹ/; however, form-focused instruction coupled with recasts helped L1 Japanese learners achieve significant gains in the production but not the perception of /ɹ/. Due to the unexpected results, Saito warned against quick conclusions and emphasized that the role of corrective feedback in L2 phonological development is still not clear. That said, overall, these recent studies suggest that corrective feedback, more specifically recasts, provided during communicative interaction may promote development of L2 phonology. However, the effects of corrective feedback on L2 phonology is still an underexplored area and more studies are needed to be able to draw confident conclusions. In an attempt to bring the issue to attention, SLA researchers have emphasized that whether interaction facilitates the acquisition of phonology is still a question waiting to be answered (Cenoz & Lecumberri, 1999; Mackey, Abbuhl, & Gass, 2012; Mackey et al., 2000). The current study is an attempt to address this call by being one of the early contributions toward filling this gap.

The main objective of this thesis is to investigate whether recasts promote the development of lexical stress placement in L2 English. It uses the following layout in order to accomplish this goal. Chapter 1, which you have just read, provides a brief history of the SLA field by focusing on the interactionist approach and highlights the need for more interactionist studies that investigate the development of L2 phonology. Chapter 2 provides operationalizations of the key SLA terms that are pertinent to the current thesis. Chapter 3 provides an overview of L2 phonology studies and makes a case for adopting an interactionist approach when studying the development of L2 phonology. Chapter 4 provides operationalizations of the key phonology and phonetics terms that are pertinent to the current thesis and explains the importance of lexical stress for second language learners. Chapter 5 discusses the pilot study and reports the results. It also explains how the pilot study informed the changes that were made for the main study. Chapter 6 and Chapter 7 discuss the methodology adopted for the main study and the results of the analyses. Finally, Chapter 8 presents a discussion of the findings by focusing on implications, limitations, and future directions.

CHAPTER 2: THE INTERACTION APPROACH

The interaction approach is a multi-dimensional approach that allows researchers to study L2 acquisition from a communicative and meaning-based perspective. The studies that adopt the interaction approach examine the relationship between L2 development and various aspects of interaction that focuses on meaning. These interactions can occur in different modes (e.g., oral, text-based) and through different mediums (e.g., face-to-face, computer-mediated). Gass (2003) explains that the interaction approach "takes as its standing point the assumption that language learning is stimulated by communicative pressure, and examines the relationship between communication and acquisition and the mechanisms (e.g., noticing, attention) that mediate between them" (p. 224). SLA researchers began to consider conversational interaction as a potential venue for learning upon the introduction of two influential positions. The first position was that communicative interactions are not merely an opportunity to practice what has been learned but they are also an opportunity for new learning to occur (Hatch, 1978; Long, 1981). In Hatch's (1978) words "one learns how to do conversations, one learns how to interact verbally, and out of this interaction syntactic structures are developed" (p. 404). The second position was introduced by the seminal works of Ferguson (1981) and Long (1981, 1983) who argued that verbal communication that takes place between native speakers and non-native speakers, particularly non-proficient ones, is different from the verbal communication that takes place between two adult native speakers. Their work introduced the concept of *foreigner talk* and drew attention to the syntactic, grammatical, and lexical simplifications in native speakers' language during communication with non-native speakers. In addition, Long (1983) observed instances of conversational adjustments by native speakers in the form of comprehension checks and clarification requests. In Long's (1983) words, comprehension checks help native speakers "assess whether they are succeeding in communicating with [non-native speakers]" and confirmation checks "tell them whether they are understanding what the [non-native speakers] are trying to communicate to them" (p. 182). This seminal body of work allowed SLA scholars to consider conversational interaction as a venue that promotes L2 acquisition. After more than three decades that have yielded research findings and numerous meta-analyses in support of the interaction approach, it is accepted that there is a robust relationship between interaction and L2 learning (Gass & Mackey, 2015). This strong relationship has been established upon careful examination of hypotheses developed around the key components that form the interaction approach, namely, input, output, noticing, attention, and corrective feedback, which facilitate L2 development in different ways.

2.1. Input

In its most common form as used in the field of SLA, the term input refers to language that is available to learners in auditory or visual form in naturalistic as well as instructional settings. Smith (1993) defines input as "potentially processible language data which are made available, by chance or by design, to the language learner" (p. 167). The role of input in facilitating language acquisition is fundamental for theories of both L1 and L2 acquisition. It is commonly agreed that input is absolutely necessary for any form of language learning to occur. In the case of L1 acquisition, the input received by infants and children in the form of spoken language is generally provided by their parents and caretakers. As children grow up and learn how to read in their native language, they also become capable of receiving input in written form. In contrast, in the case of L2 acquisition the form and the type of input that learners receive or are exposed to vary greatly depending on various factors such as age, socio-economic factors (e.g., a university student vs. an immigrant worker), and the context (e.g., foreign-language context vs. second-language context). As the focus of the current study is L2 development, this section will briefly explain the role of input from an SLA perspective.

Krashen's (1983, 1985) proposal of the monitor model, affective filter, and input hypothesis was highly influential in shaping the research agenda during the early years of SLA. Krashen defined comprehensible input, also referred to as *i+1*, as language input that is slightly beyond the learner's current level of linguistic competence. He argued that the only way for L2 learners to acquire new target language forms would be through exposure to comprehensible input when they have low affective filter, which he defined as a mental block that prevents comprehensible input from being processed by the learner's cognitive mechanism. His views were aligned with the views of L1 researchers who argued that children learn their native language from positive evidence alone (e.g., Braine, 1971; Marcus, 1993; Pinker, 1989), that is, well-formed models of the target language. As comprehensible input is a form of positive evidence, Krashen's proposition that comprehensible input is sufficient for successful language acquisition also meant that negative evidence, which is feedback by a more competent speaker on the learner's non-target-like production, does not make a major contribution to L2 acquisition. The usefulness of comprehensible input was accepted by other SLA researchers; however, there was also criticism in terms of a) what *i+1* really means, and b) the role of comprehensible input as the only condition necessary for language learning.

A major issue surrounding the concept of *i+1* was that it is not possible to define it clearly or measure it objectively, which are necessary from the perspective of the scientific method. Another important issue was the assumption that all input was automatically processed by the learner. Corder (1967, 1982) addressed this problem by distinguishing between input and intake. He defined input as "what is available for going in" and intake as "what goes in" (Corder, 1967, p. 165). In a similar vein, Gass (1988) proposed a model for L2 acquisition that incorporated two types of input: *apperceived input* and *comprehended input.* She defined apperceived input as what learners notice in the ambient speech, highlighting that not all input is noticed by learners. According to Gass's model, apperceived input precedes comprehended input, which she defined as the amount of apperceived input that learners comprehend. Gass also distinguished comprehended input from comprehensible input explaining that the former implies that the learner is in charge of the act of comprehending as opposed to the latter which disregards learner control over input. Another difference between these two types of input is that, unlike comprehensible input, comprehended input is not a dichotomous term. According to Gass's model, input can be comprehended in a number of different ways as in overall utterance meaning or at a metalinguistic level such as phonological or syntactic features.

The other issue with Krashen's model was the lack of emphasis on the role of negative evidence in L2 acquisition. As one of the earliest critics of Krashen's proposition, Schachter (1984) highlighted the importance of negative input, which she defined as "information provided to the learner that her utterance was in some way deviant or unacceptable by the native speaker" (p. 168), as a fundamental element that aids L2 acquisition. Schachter (1986) argued that negative input is not limited to corrections and could include clarification requests and confirmation checks. In other words, she considered any type of communicative move that highlights a problem with the learner's production as negative input, which was later referred to as negative evidence. The potential of negative evidence in promoting L2 acquisition immediately became a topic that drew researchers' interest (e.g., Brock, Crookes, Day, & Long, 1986; L. White, 1991; Long, 1990; Varonis & Gass, 1985;), and as the SLA field continued to evolve, negative evidence has become a core construct that led to the birth and investigation of other core constructs such as attention, noticing, and various types of corrective feedback.

In sum, language input in SLA may refer to positive evidence or negative evidence, the first being error-free models of the target language and the latter any indication that the learner's production was erroneous. Chapter 2.5 discusses the types of negative evidence in more detail.

2.2. Output

Another complement to Krashen's comprehensible input hypothesis was Swain's (1985, 1995) comprehensible output hypothesis. Upon studying the language development of learners in French immersion settings at Canadian schools, a context known for abundant opportunities for receiving comprehensible input, Swain (1985) found that learners were unable to develop in terms of grammatical accuracy. Her findings contradicted Krashen's (1985, 1994) position that comprehensible input was the driving force behind L2 acquisition. Swain suggested that comprehensible output, in other words learners' attempt at producing the target language, is also necessary for L2 acquisition. Swain (1995) claimed that output allows noticing of the gap between "what [learners] *want* to say and what they *can* say" (emphasis in original, p. 126). In this sense, the comprehensible output hypothesis draws on the noticing hypothesis (Schmidt, 1990; Schmidt & Frota, 1986), and as such, it highlights the role of consciousness as a crucial component for L2 acquisition. Swain (1995) explained that output facilitates L2 acquisition through its three major functions: it allows learners to test their hypotheses related to syntactic features, notice non-target-like elements in their own production, and reflect on those elements which would potentially lead to the unlearning of errors. A number of empirical studies have provided evidence for these three functions (e.g., Adams, 2003; Izumi, 2002; Izumi & Bigelow, 2000; Mackey, 2002; Philp & Iwashita, 2013; Swain & Lapkin, 1998). In her later works, Swain (2005) also argued that output promotes language fluency.

In its basic form, the comprehensible output hypothesis is mainly concerned with production of language and it posits that output that is produced by the learner for communicative purposes, not necessarily as a response to a clarification request, will create opportunities for noticing the gap in their knowledge. This claim has been supported by research findings (e.g., Philp & Iwashita, 2013; Swain & Lapkin, 1995). Other scholars such as Pica, Holliday, Lewis, and Morgenthaler (1989), however, argued that "especially helpful to the acquisition process is the interaction in which learners are 'pushed' to make their output comprehensible" (p. 64.). In this context, pushed output refers to learners' self-modification of an utterance in response to a signal from their interlocutor indicating that the output has not been fully comprehended. In a more recent discussion on the role of output, Mackey (2012) refers to pushed output as modified output and defines it as reformulation of "one's original utterance in response to feedback or self-monitoring" (p. 16). Mackey also highlights various benefits of the process that learners go through when modifying their utterance regardless of whether it is self-initiated or pushed.

In short, both output and modified output are instrumental for L2 acquisition as they provide language learners with opportunities to use the target language productively, which in return leads to hypothesis testing, receiving further input in the form of positive and negative evidence, and noticing the gap between the intended production and the actual production or the negative evidence provided by their interlocutor. As such, output hypothesis has become a main construct within the interaction approach due to its role as a trigger for a range of developmental processes (Gass & Mackey, 2006).

2.3. Noticing and attention

Noticing and attention are two core constructs within the interaction approach that underscore the role of consciousness in L2 acquisition. Schmidt's (1990, 1992, 1994, 2001) influential work on noticing is based on the premise that learners need to notice new target features in the input in order to be able to convert those target features into intake. Schmidt argued that noticing new target language features as well as the difference between those features and learners' own non-target-like production, commonly known as noticing the gap, is a condition necessary for L2 acquisition to occur. In the same way, in order to benefit from negative evidence provided in the form of corrective feedback, learners need to be aware that they are being corrected and they need to notice the difference between the negative evidence and their preceding production. In this sense, noticing is not simply a matter of detection but rather "what is both detected and then further activated following the allocation of attentional resources" (Robinson, 1995, p. 297). The roots of the noticing hypothesis go back to Schmidt and Frota's (1986) paper that presented the findings of a case study of Schmidt's own language learning experience in Brazil over a period of five months. The authors suggested that interaction, input, or classroom drills were not sufficient for Schmidt to learn particular verb forms in Portuguese and that he needed to notice these forms in the input for acquisition to occur. In this early work on noticing, the authors avoided making strong claims about whether noticing occurred consciously or unconsciously. In his later works, Schmidt (1994, 2001) argued for a clear role of consciousness in language acquisition. He identified four types of consciousness relevant to the context of L2 learning: consciousness as intentionality, which refers to whether learning happens deliberately or unintentionally, the latter leading to incidental learning; consciousness as attention referring to a mental state controlled by the channeling of attention toward a particular stimulus; consciousness as awareness, which refers to whether learning happens implicitly or explicitly; and finally consciousness as control, which makes the point that fluency in terms of language performance increases when explicit knowledge is not consciously used during language construction (Schmidt, 1994).

Attention has been a key element in the discussions that surround the noticing hypothesis. Schmidt (1993, 1994) described attention as an essential precursor to noticing that helps with converting input into intake. He also argued for an inverse relationship between attention and fluency when defining fluency as a procedural skill that requires the use of little to no attention by the learner (Schmidt, 1992). Later on, drawing upon theories of learning discussed in the field of psychology, Schmidt (2001) proposed the strong version of the noticing hypothesis with increased emphasis on the role of attention by suggesting that "attended learning is far superior, and for all practical purposes, attention is necessary for all aspects of L2 learning" (p. 3). Although the strong version of Schmidt's noticing hypothesis received criticism (e.g., Gass, 2018), the role of attention and noticing are considered two essential factors that are necessary for L2 development.

An alternative view of the role of attention in SLA was proposed by Tomlin and Villa (1994). In their seminal paper, they treated attention as a construct that captures three main processes: alertness, orientation, and detection. In their model, alertness refers to overall readiness for the incoming stimuli, in order words, input; orientation is the channeling of attention to a specific type of input; and lastly, detection is the actual mental registration of the input that was attended to, which is similar to noticing in Schmidt's model. However, a major difference between the two models is that for Schmidt (2001) noticing and awareness are directly linked (e.g., "noticing [i.e., becoming aware of] the structural regularities of a language" p. 5), whereas for Tomlin and Villa (1994) detection "does not require awareness" (p. 197).

Apperception is another construct that needs to be mentioned due to its connection to Schmidt's noticing hypothesis. Gass (1988) introduced a five-step model for L2 acquisition that explains the processes involved in converting input to output, and apperception plays a significant role in this model (see Gass, 2018 for the most recent version of the model). Gass defines apperception as the portion of input that learners notice in ambient speech. She explains that a crucial difference between noticing and apperception is that apperception is driven by past experiences, which could include knowledge of the native language, the existing knowledge of the target language, world knowledge, and so on (Gass, 2018). This means that a learner can only apperceive language stimulus as long as he or she can make connections between that stimulus and a past experience or current knowledge. Apperception encompasses both attention and awareness, which distinguishes Gass's model from Tomlin and Villa's model in terms of the role of awareness. Gass's model is also different from Schmidt's model in terms of the processes involved in converting input to intake. According to the noticing hypothesis, what is noticed in the input becomes intake; whereas apperceived input needs to be comprehended in order to become intake. Comprehended input, different from Krashen's comprehensible input, places an emphasis on the learner's cognitive mechanism rather than simply defining an aspect of the stimulus.

In short, regardless of the conceptual differences and the debate on the level of detail attention models in SLA should account for (for a detailed discussion see Leow, 2002; Robinson, 2003), there is a consensus in the SLA field supporting the view that attention to and noticing of target language features in the input are necessary conditions for L2 development.

2.4. Salience

The established role of noticing and attention in SLA postulates that factors that facilitate noticing and draw learners' attention to a target language feature would naturally promote L2 development. In this sense, the salience of a linguistic feature becomes a factor that could potentially guide attention toward that particular feature, which would then lead to the possibility of that feature to be noticed. The reasoning behind this conclusion is that salient language features will be attended to, and as a result they will be noticed, and then processed by the learner (N. Ellis, 2016). According to *Merriam-Webster Dictionary*, one of the definitions of *salient* (n.d.) is "standing out conspicuously." Although theoretical definitions of salience are more detailed and varied in comparison to a brief dictionary entry, for the most part, the operationalizations used by interactionist researchers are built around the idea that salient forms stand out in the input due to various structural characteristics that render them more prominent than the surrounding input. For example, according to Loewen and Reinders (2011), "salience refers to how noticeable or explicit a linguistic structure is in the input" (p. 152). This type of salience is also referred to as *perceptual salience* as it involves the action of noticing by the learner. In a detailed discussion on salience from a usage-base perspective, N. Ellis (2016, 2018) draws on psychological research and explains that there are three types of salience, namely, *psychophysical salience, salient associations,* and salience due to the effect of *context and surprisal*. He explains that psychophysical salience refers to inherent qualities of the stimulus, such as being brighter or louder, that make it stand out in contrast to surrounding stimuli. Salient associations, on the other hand, are memory or expectation driven. In other words, prior knowledge of or experience with a particular stimulus can make it stand out among other stimuli during subsequent encounters. Finally, N. Ellis explains that in the case of salience that stems from context and surprisal, a feature of the stimulus that deviates from the ordinary or probable, which is calculated by the learner's cognitive mechanism, has the potential to catch the learner's attention and therefore be processed. In summary, the first category that N. Ellis presents is mainly related to physical aspects of the stimulus whereas the other two are shaped by what the learner already knows about language and the world. N. Ellis (2016) refers to the interplay among the stimulus, the context, and the knowledge of the learner to emphasize the complexity of salience as a construct, which makes it difficult to define. Similarly, many SLA scholars have expressed that salience is a construct that evades an easy definition (Leeman, 2003; Mackey, 2012; VanPatten & Benati, 2015). That said, for practical purposes and due to its relevance in the context of the current study, throughout the rest this dissertation the terms *salience* and *salient* will be used to refer to what N. Ellis (2016, 2018) defines as psychophysical salience.

Linguistic input can be salient due to occurring at high frequency (N. Ellis, 2002) or low frequency (Rácz, 2012). During the early stages of acquisition, a target feature that occurs at high frequency in the input may be salient and therefore noticed by the learner; however, as the learner advances and establishes competence in their L2, a feature that is less frequent may become novel or salient, and as a result, catch the attention of the learner (Gass, 2018). Phonetic characteristics of the input such as syllabicity can also impact its salience (VanPatten & Benati, 2015). An earlier definition of salience by Dulay, Burt, and Krashen (1982) highlight phonetic substance, in particular stress, as an element that contributes to salience of a linguistic item:

> Psycholinguistics have defined salience by referring to particular characteristics that seem to make an item more visually or auditorily prominent than another. Such characteristics include the amount of phonetic substance (whether the item is a syllable or not); or the stress levels of an item (the amount of emphasis placed on it). (p. 33)

Similarly, Gass (2018) mentioned that “although the learner has to perceive salience, actual salience can be affected by such things as stress and loudness” (p. 57). It has been argued that phonetic or phonological salience can impact acquisition of certain morphological features or function words. For example, N. Ellis (2006) explains that morphological features that have low phonological salience such as the third person singular -*s* or function words that do not carry stress are more difficult for L2 learners to acquire. N. Ellis refers to a study by Herron and Bates which showed that adult native speakers had a success rate of 50% or less when identifying function words that were cropped from connected speech and presented in isolation (as cited in N. Ellis, 2006, p. 171). Expanding onto this piece of evidence, which shows that lack of phonetic salience can make it difficult for native speakers to identify function words, N. Ellis argued that the lack of salience is one of the reasons why L2 learners struggle with hearing function words in natural speech and learning them.

Language features can have high psychophysical salience due to inherent qualities such as the nature of phonetic structure as discussed in the example above. Alternatively, certain features of the input can be made more salient compared to surrounding input through external manipulation. This type of manipulation is referred to as input enhancement (Smith, 1991, 1993). In its most basic form, input enhancement is used by language teachers frequently when they circle or underline a particular word on the whiteboard/blackboard or when they pronounce a particular word or morpheme with extra prosodic emphasis in order to draw learners' attention to the particular language feature they are trying to teach. When input enhancement occurs in the form of highlighting, underlining, bolding, or using different a font or font size, it is called textual enhancement (Izumi, 2002; J. White, 1998; Leow & Martin, 2018). In the case of auditory input, additional emphasis through stress and intonation can be used to increase the salience of a particular feature (Leeman, 2003; Loewen & Philp, 2006). In addition to textual and auditory enhancement of the input, corrective feedback can also be used to increase perceptual salience of target language features to facilitate their acquisition (Leow, 2015; Long & Robinson, 1998). Smith (1993) calls this *negative input enhancement* and explicates that it "would flag given forms as incorrect, thus signaling to the learner that they have violated the target norms" (p. 177). For example, as a commonly used form of corrective feedback, recasts have the potential to make the correct form of attempted target features more salient due to their contingency to the learner's erroneous production which gives the learner the opportunity to juxtapose their production to the reformulation provided in the form of a recast (Long, 2007). At the same time, recasts themselves can be made more salient through the use of emphatic stress (Leeman, 2003; Loewen & Philp, 2006). The salience and noticing of recasts can also depend on the linguistic feature targeted with studies showing that learners notice recasts that target phonological features (Carpenter, Jeon, MacGregor, & Mackey, 2006; Mackey et al., 2000). In short, there can be a myriad of factors that determine the level of salience a linguistic feature or even a particular form of corrective feedback can have. That said, in the case of recasts, both prosodic emphasis and contingency play an important role.

2.5. Corrective feedback

Linguistic input comes in two forms: positive evidence and negative evidence. Long (1996) notes that positive evidence is “models of what is grammatical and acceptable (not necessarily the same)” (p. 413). In other words, positive evidence refers to exemplars of the target language produced by native or proficient non-native speakers that non-proficient L2 speakers hear or read. Negative evidence, on the other hand, is information about what is not grammatical or acceptable in the target language. In many interactionist SLA studies, negative evidence is operationalized as feedback on learner errors. However, since the word feedback is a neutral term in essence and because it can be provided as a validation of target-like production as well, SLA researchers use the term *corrective feedback* to distinguish feedback that is provided on learner errors from positive feedback that functions as a validation of problem-free production. Furthermore, what many interactionist researchers particularly focus on is “reactive feedback — that is, feedback that occurs as a reaction to some linguistic problem” (Gass & Mackey, 2006, p.7). In the case of L1 acquisition, children may receive corrective feedback from their parents when their language production is non-target-like. Similarly, adult learners of a second language may receive corrective feedback as part of classroom instruction or during everyday interaction with native speakers or other non-native speakers. Researchers of child language acquisition have been skeptical about the usefulness of negative evidence, and hence corrective feedback, and argued that there is no proof or data to suggest that negative evidence is a necessary condition for children to learn their L1 (e.g., Marcus 1993; Pinker 1989). However, the general view in the SLA field is that L1 acquisition is different from L2 acquisition because in the case of L2 acquisition learners sometimes make incorrect generalizations based on their L1, which necessitates the provision of corrective feedback to indicate what is structurally possible in their L1 may not be possible in their L2 (L. White, 1991). Therefore, SLA scholars view corrective feedback as a facilitative component in L2 acquisition (for a recent review, see Yilmaz, 2016). Moreover, SLA studies have shown that corrective feedback is a commonly occurring phenomenon in classroom settings (Loewen, 2009; Lyster & Ranta, 1997; Oliver, 2000; Yoshida, 2008) and during interactions between children or adult speaker dyads as well as between a native speaker and a non-native or between two non-native speakers (Mackey, Oliver, & Leeman, 2003; Oliver, 2009). It may be provided by teachers or more competent students in classroom contexts, and by native-speakers or more competent L2 speakers during laboratory research or in naturalistic contexts. SLA scholars have argued that corrective feedback that draws learners' attention to their non-target-like production is necessary for successful L2 acquisition (Gass, Mackey, & Pica, 1998; Lightbown & Spada, 1990; Long, 1990, 1996; Schachter, 1986, 1991; Schmidt & Frota, 1986). The rationale behind this

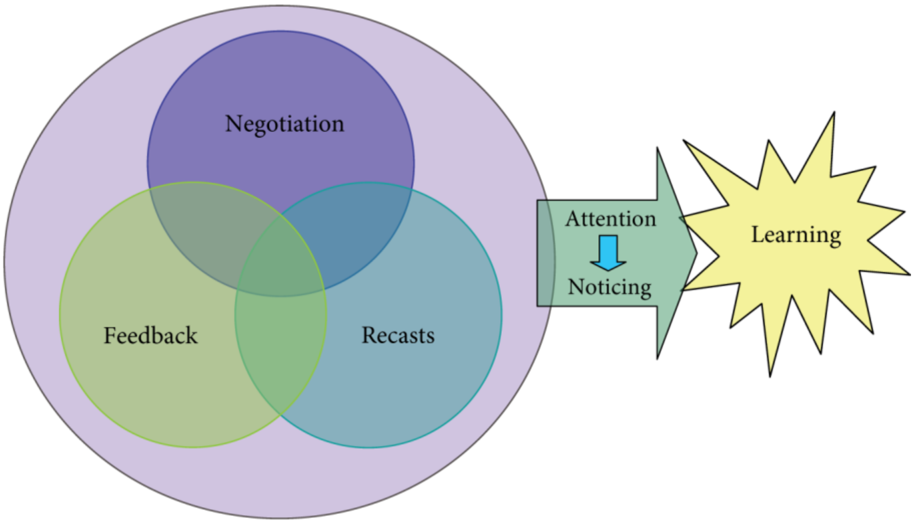


Figure 2.1 Model of interaction and learning (Reprinted from Gass & Mackey, 2006)

argument is that attention and noticing are two key elements that bridge the gap between learner-internal and learner-external factors to facilitate L2 development (Schmidt, 2001) and corrective feedback that signals a problem in the learner's output has the potential to trigger these two fundamental cognitive functions (see Figure 2.1). In fact, research studies have highlighted the role of attention as an agent that may allow learners to notice the illocutionary force of corrective feedback (Goo & Mackey, 2013; Izumi & Bigelow, 2000; Li, 2010; Mackey, 2006).

The interactionist literature has placed different types of corrective feedback into two main categories called implicit and explicit feedback (R. Ellis, 2007, 2008; Leeman, 2007; Loewen, & Erlam, 2006; Mackey, 2012). Implicit feedback does not directly or openly point at the erroneous output by the learner but rather signals it covertly. On the other hand, explicit feedback is an attempt to directly and overtly draw the learner's attention to the error. Mackey (2012) explains that although the main function of both types of feedback is to signal a problem with learners' non-target-like production, "explicit feedback (i.e. indicating a problem with the utterance with a focus on the form) can interrupt the flow of the interaction, whereas implicit feedback (i.e. indicating there is a problem with primary focus on meaning) usually allows for the interaction to continue uninterrupted" (p. 116). Beyond these two main categories, corrective feedback can be categorized depending on the specific function that each feedback type serves. Lyster and Ranta's (1997) seminal work identified six main types of corrective feedback moves that fall under the implicit-explicit continuum. These feedback moves are explicit correction, metalinguistic feedback, elicitation, clarification requests, repetition, and recasts. The first three of these six types of feedback moves are considered explicit feedback and the last three implicit feedback.

Explicit correction provides the correct form with an overt indication that the learner's production was incorrect:

S: One upon a time

T: No we say once

S: Once (Oliver & Mackey, 2003, p.524)

Metalinguistic feedback provides comments on what is wrong in the learner's production without providing the correct form:

Learner: He kiss her

Researcher: Kiss—you need past tense

Learner: He kissed (Ellis et al., 2006, p. 353)

Elicitation is a form of corrective feedback that aims to push the learner to provide the correct form by using a question or by pausing before the target form and allowing the learner to fill in the blanks:

Student: …on the street there was a policeman, and she was skipping running.

Teacher: I am sorry, she was…?

Student: Skipping running, the thief (Nassaji, 2009, p.429)

Clarification requests are a type of feedback that indicates that the learner's production was ill-formed or the interlocutor did not understand the message:

Learner: Why does he taking the flowers?

Researcher: Pardon?

Learner: Why does he take flowers? (Loewen & Nabei, 2007, p.367)

Repetition is a type of feedback that repeats the learner's error without any changes but highlights the error using stress and intonation:

Student: La chocolat . . . "(F) Chocolate."

Teacher: La chocolat? "(F) Chocolate?"

Student: Le chocolat. "(M) Chocolate." (Lyster, 2004, p.405)

Last but not least, recasts are a type of feedback that provides a target-like reformulation of the learner's non-target-like production:

Learner: And I saw a boy next to the bar. I think he was with his girlfriend. They talking to each other.

Researcher: They were talking to each other. (Révész, 2012, p. 94)

These six corrective feedback moves can also be grouped based on whether they provide the correction to the learner or request self-repair by pushing the learner to correct their own error. Lyster (2004) defined the latter as *prompts* and explained that "they withhold correct forms (and other signs of approval) and instead offer learners an opportunity to self-repair by generating their own modified response" (p. 405). In subsequent publications, the first type of feedback has been also been referred to as *input-providing* and the latter *output-prompting* (R. Ellis, 2008; Sheen & Ellis, 2011). Loewen and Nabei (2007) offered a categorization of the six types of corrective feedback moves based on whether they are input-providing or output-prompting mapped on an implicit-explicit continuum. According to their categorization, explicit correction and recasts are both input-providing feedback moves with explicit correction being the explicit type and recasts being the implicit type; metalinguistic feedback, elicitations, repetitions, and clarification requests are the output-prompting moves, with metalinguistic feedback on the explicit end of the continuum and clarification requests on the implicit end.

Meta-analyses so far have indicated a positive effect of corrective feedback on language development (Goo, Granena, Yilmaz, & Novella, 2015; Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007; Russell & Spada, 2006). A great amount of SLA research has attempted to tease apart the overall effectiveness of different corrective feedback moves and answer the question of whether implicit feedback is better than explicit feedback or vice versa. While some researchers have argued that explicit feedback may be more beneficial than implicit feedback (Ellis et al., 2006; Lyster, 1998a; Lyster & Ranta, 1997; Lyster & Saito, 2010; Sheen, 2007), others have argued that implicit feedback, particularly recasts, are more effective (e.g., Long, Inegaki, & Ortega, 1998; Long, 2007). There have been also studies that found no significant difference between explicit and implicit feedback in terms of their level of effectiveness (Goo, 2012; Loewen & Nabei, 2007; Sauro, 2009). However, it is difficult to argue that one type of feedback has overall advantage compared to the other because the effectiveness of a particular type of corrective feedback depends on various factors such as target language feature (Mackey et al., 2000; Saito & Lyster, 2012a); outcome measures (Révész, 2012); task complexity (Baralt, 2013; Révész et al., 2014); and learner factors such as proficiency level (Ammar & Spada, 2006), developmental readiness (Mackey & Philp, 1998), working memory capacity or language analytic ability (Yilmaz, 2013). As such, due to the amount of methodological variability among corrective feedback studies, caution should be exercised when coming to conclusions about which type of feedback is better than the other. It is also important to note that meta-analysis studies have shown that the effect of explicit feedback can be observed immediately, while the effect of implicit feedback is usually delayed (e.g., Li, 2010; Mackey & Goo, 2007). Therefore, these meta-analyses warned against hasty conclusions and comparisons between implicit and explicit feedback. Due to the existence of a multitude of factors that mediate the effectiveness of corrective feedback making the interpretation of comparative studies difficult, Goo and Mackey (2013) suggested that it would be more fruitful to direct research efforts toward investigating how various types of corrective feedback impact different target-language features in different contexts.

2.6. Recasts

Recasts are one of the most extensively researched form of corrective feedback, probably because they are a common type of corrective feedback that occurs in both L1 and L2 acquisition contexts. Despite some mixed findings on the effectiveness of recasts in L2 acquisition (e.g., Lyster, 2004; Lyster & Ranta, 2013), a sizable number of studies have found recasts to be useful for L1 acquisition (e.g., Chouinard & Clark, 2003; Farrar, 1992; Saxton, 1997, 2005) as well as L2 acquisition (e.g., Goo & Mackey, 2013; Mackey et al., 2003; Mackey & Philp, 1998; McDonough & Mackey, 2006; Loewen & Philp, 2006; Philp, 2003; Saito, 2013a; Saito & Lyster, 2012a). As an extensively researched type of corrective feedback, recasts have been defined by various scholars. Below are some of the definitions by interactionist scholars who are well-known for their work on corrective feedback:

> A *corrective recast* may be defined as a reformulation of all or part of a learner's immediately preceding utterance in which one or more non-target like (lexical, grammatical, etc.) items is/are replaced by the corresponding target language form(s), and where, throughout the exchange, the focus of the interlocutors is on *meaning*, not language as an object. (emphasis in original, Long, 2007, p. 77)

> A response to an error in learners' oral production that involves the reformulation of the incorrect linguistic element while maintaining the overall meaning of the utterance. (Loewen & Reinders, 2011, p.148)

> …teacher's reformulation of all or part of a student's utterance, minus the error. (Lyster & Ranta, 1997, p. 46)

> …a form of feedback in which learners are provided with more target-like versions of their immediately preceding non-target-like utterance. (Mackey, 2012, p.14)

As can be seen from the definitions above, there are two crucial characteristics of recasts: they are provided immediately after a learner error and they are meaning-focused. To begin with the emphasis on the focus on meaning, it does not necessarily mean that recasts are unnecessary if the interlocutor can extract the meaning from the learner's non-target-like utterance. It means that recasts serve a communicative function and are less likely to break the flow of conversation unlike some other types of corrective feedback that will inevitably disrupt the natural flow of communication. In fact, teachers utilize recasts frequently and during research studies SLA researchers provide recasts on learners' erroneous utterances despite being able to understand the actual content of the message. This type of recasts is called *didactic recasts* defined as "reformulations of erroneous utterances that occur even in the absence of a communicative problem" (R. Ellis, 2017, p.18). SLA scholars who argue that recasts can be implicit or explicit depending on the context consider didactic recasts as an explicit form of corrective feedback (e.g., Ellis & Sheen, 2006; Sheen & Ellis, 2011). Recasts, including the ones that are didactic, are meaning-focused in that they do not interrupt the natural flow of communication with metalinguistic explanations or by requiring the learner to fix his or her erroneous production before continuing with the conversation. This is one of the main advantages of recasts as a form of corrective feedback because they allow learners to focus on meaning and communication while providing them with negative evidence. When it comes to immediacy of recasts, it allows learners to juxtapose their ill-formed production with the target-like production provided by the interlocutor allowing cognitive comparison of two contrasting forms (Doughty, 2001). This characteristic of recasts has been discussed as part of the *contrast theory* in L1 acquisition literature (e.g., Saxton, 1997) and it has been referred to as *contingency* in L2 acquisition literature (e.g., Long, 2007).

Recasts are multifaceted, and as a result, they can come in various forms. There have been attempts at identifying different characteristics of recasts in order to inform methodological decisions and to allow meaningful comparisons across studies. Lyster (1998a) introduced a four-way categorization that he used for coding recasts: an isolated declarative recast, which is a correct reformulation of "all or part of the utterance with falling intonation and no additional meaning;" an isolated interrogative

Table 2.1

Single-move recasts

Characteristic	Sub-category	Description
Mode	Declarative	Provided in a statement
	Interrogative	Provided in interrogative form
Scope	Isolated	Does not include additional information
	Incorporated	Includes additional semantic content
Reduction	Reduction	Shorter than the erroneous utterance
	Non-reduction	Reformulation of the complete erroneous utterance
Length	Word/short phrase	Only one word or phrase with one content word
	Long phrase	More than two words excluding a finite verb
	Clause	At least two phrasal constituents with a finite verb
Number of changes	One change	Change of only one linguistic item
	Multiple changes	Change of multiple linguistic items
Type of change	Addition	Suppliance of missing grammatical element
	Deletion	Removal of a linguistic element
	Substitution	Replacing of one linguistic element with another
	Reordering	Change of order of elements
	Combination	A combination of different types of change
Linguistic focus	Pronunciation	Reformulation targets pronunciation
	Vocabulary	Reformulation targets vocabulary
	Grammar	Reformulation targets grammar

Adapted from Sheen (2006)

Table 2.2

Multi-move recasts

Characteristic	Description
Corrective recasts	Preceded by repetition of the error
Repeated recasts	Partial or complete repetition of recasts
Combination recasts	Occur with other corrective feedback types except explicit correction

Adapted from Sheen (2006)

recast, which is a correct reformulation of "all or part of the utterance with raising intonation and no additional meaning;" an incorporated declarative recast, which is a "correct reformulation of all or part of a learner's utterance into a longer statement"; and finally, incorporated interrogative recast, which is a "correct reformulation of all or part of a learner's utterance into a question" (pp. 58-59). Later, Philp (2003) identified the length of a recast as well as the number of changes in a recast as two key characteristics that have direct implications in terms of working memory load and learners' attention sources, hence learners' ability to notice the corrective force of recasts. Drawing on these research findings, Sheen (2006) proposed a rather detailed taxonomy based on various unique characteristics that recasts have, which is summarized in Table 2.1 and Table 2.2.

According to the taxonomy provided by Sheen (2006), the recast on the next page is a single-move recast that is declarative, isolated, and reduced. It consists of a single-word with one-change and the type of change is substitution. The linguistic focus of the recast is pronunciation.

Student: It’s good to start reading /ɾidɪŋ/ something.

Teacher: Reading /ɹidɪŋ/.

Student: Reading /ɹidɪŋ/ something. Even it is comic books or novels.

(Saito, 2015b, p.277)

The taxonomy provided by Sheen (2006) encourages researchers to consider various aspects of recasts, which is helpful for operationalization of recasts at the study design stage as well as coding and analysis of data in the subsequent stages. It also shows that recasts are not a single form of feedback but come in multitude of forms. The multifaceted nature of recasts also has implications in terms of the implicitness and explicitness debate. Although recasts have been identified as a form of implicit feedback in the past (e.g., Long, 2007; Lyster, 1998a), many SLA scholars have argued that recasts are rather on an implicit-explicit continuum as some types of recasts are easier to notice (Ellis & Sheen, 2006; Erlam & Loewen, 2010; Loewen & Philp, 2006; Lyster, Saito, & Sato, 2013; R. Ellis, 2008).

An important issue concerning recasts has been how they measure up compared to other types of corrective feedback and whether they are useful for learners or not. Despite a number of studies that found recasts to be less effective than other types of corrective feedback, particularly metalinguistic feedback, and the following discussions that are skeptical about the usefulness of recasts (e.g., Ammar & Spada, 2006; Ellis et al., 2006; Lyster, 1998a; Lyster & Ranta, 2013; Lyster & Saito, 2010), there are also studies that did not find a significant difference between metalinguistic feedback and recasts in terms of their effectiveness (Goo, 2012; Loewen & Erlam, 2006; Loewen & Nabei, 2007; Sauro, 2009). What is more, there is already substantial evidence indicating that recasts can be facilitative of L2 development (Egi, 2007; Goo & Mackey, 2013; Ishida, 2004; Iwashita, 2003; Mackey & Goo, 2007; Li, 2010, 2014; Long, 2007). An extensive number of studies found a positive relationship between recasts and a wide range of target language features such as question formation (Loewen & Nabei, 2007; Mackey & Philp, 1998; Philp, 2003), simple past tense (Doughty & Valera, 1998; McDonough, 2007), past progressive tense (Révész & Han, 2006), possessive determiners (Ammar & Spada, 2006), that-trace filter (Goo, 2012), grammatical gender (Lyster & Izquierdo, 2009), adjective ordering and locative construction (Long et al., 1998), noun-adjective agreement (Leeman, 2003), and development of the approximant /ɹ/ (Saito, 2013a, 2015a; Saito & Lyster, 2012a) among many others. Empirical studies have also highlighted various characteristics of recasts that increase their efficacy such as freeing up processing load (Ammar & Spada, 2006), facilitating language development for learners who are developmentally ready (Han, 2002; Li, 2014; Mackey & Philp, 1998; Trofimovich, Ammar, & Gatbonton, 2007), and having long-term durable effects (Mackey & Goo, 2007; Li, 2010). Recasts can also be a source of positive and negative evidence simultaneously (Leeman, 2003), which provides learners with an opportunity to juxtapose their erroneous production with the correct form as they are generally provided immediately after an error has occurred (Mackey, 2007; Saxton, 2005). With all these benefits, recasts are in fact one of the most commonly occurring feedback moves in natural communicative settings between native speakers and non-native speakers whether the dyads are children or adult (Braidi, 2002; Oliver, 1995, 2009), as well as in language classrooms as part of the communication between teachers and learners (Brown, 2016; Lyster & Ranta, 1997; Panova & Lyster, 2002, Sheen, 2004).

CHAPTER 3: PHONOLOGY AND SLA

In first language acquisition, speech precedes writing. Children learn how to speak before they learn how to write. Apart from the case of children, oral communication has an important role in daily interactions between people of all ages whether it is a face-to-face conversation, or an exchange over the phone or through a video-conferencing software. Despite these facts and the obvious communicative value of pronunciation, research that focuses on L2 pronunciation and phonological development has received relatively less attention in comparison to other areas of SLA research (Derwing & Munro, 2005). According to Foote and Trofimovich (2018), one of the possible reasons for marginalization of pronunciation research is the lack of theoretical framework and models that pronunciation research can build on. They also emphasize that existing SLA theories are rarely applied to pronunciation research. What is more, Derwing and Munro (2015) point out that most research on L2 speech does not focus on the impact of instruction on development, and as a result, does not provide implications for the language classroom. Other reasons for the lack of interest in pronunciation research had been discussed earlier in Leather's (1983) seminal review on L2 pronunciation studies. Leather explained that the problem of setting the wrong goals for L2 pronunciation, that is aiming for native-like production, had led to research endeavors being perceived as unproductive use of time. Another reason for the minimal interest in teaching pronunciation and the skepticism about its teachability was the dissatisfaction with a process that was perceived to be meaningless due to a lack of communicative value (e.g., drill exercises) and the low success rate among students (for a review see Morley, 1991). This was particularly the case when the SLA field was under the influence of behaviorism. Approaches such as the direct method and the audio-lingual method mainly focused on drills with the unrealistic goal of helping learners achieve native-like pronunciation and they were far from being meaning-based activities. Following this period was the rise of communicative language teaching partially motivated by Krashen's (1977, 1985) comprehensible input hypothesis. Following Krashen's views on language acquisition, it was thought that learners would be able to improve their pronunciation through the comprehensible input they receive during communicative activities. Also, during this period, the goal of teaching was to help learners improve their communicative competence (Hymes, 1972). Therefore, unless absolutely necessary such as in the case of a communication breakdown, feedback on pronunciation was seen as a hindrance that prevents learners from improving their fluency and communicative competence. Consequently, there was a period of time during which teachers either ignored pronunciation or did not know how to deal with it. However, research on pronunciation started to evolve in accordance with the paradigm shift that placed importance on intelligibility and comprehensibility (Derwing & Munro, 1997; Levis, 2005) as a useful goal instead of native-like production. Intelligibility as a construct refers to how much the listener understands a particular speaker, usually measured via a transcription task. Comprehensibility as a construct refers to listeners' perceptual judgment of how much effort was required on their part in order to understand the speaker, usually measured on a 7-point or 9-point likert scale. Despite the shifting focus from native-like speech to intelligibility and comprehensibility, research on pronunciation as well as pronunciation instruction has been considered relatively inessential compared to other linguistic or communicative skills required for successful L2 acquisition. Isaacs (2009), and Levis and Sonsaat (2018) provide a succinct historical account of events that shed light on the ups and downs that pronunciation research has gone through since the 1960s. Despite the paradigm shift and the recent increase in interest in phonological development in L2, the research on the acquisition of L2 pronunciation still falls behind when compared to SLA research that focuses on morphology or syntax. This is particularly true in the case of SLA studies that adopt the interaction approach to language development.

This does not mean that there are no studies on L2 pronunciation. Some studies are theory-driven and their main goal is to test hypotheses related to non-native speech directed at examining various factors such as the relationship between perception and production or the effect of age on native-like speech production. However, even when these studies are considered, there is an underrepresentation of L2 phonology studies within the broader SLA field. Major (1998) points out that only a limited number of studies published in reputable journals explore questions related to L2 phonology, and in general, they are relatively limited in scope. It could be argued that the issue of limited scope is due to the particular focus on segmental features. A decade after Major's observation, Gut (2009) surveyed 16 leading linguistics journals published between 1969 and September 2008 to analyze representation and scope of L2 phonology studies in those journals. Her findings showed that out of the 176 phonology studies surveyed only 24 studies focused on suprasegmentals (10 studies on word stress, 9 studies on intonation, 4 studies on speech rhythm, 1 study on prosody). Although Gut's study was conducted about a decade ago and some progress has been achieved since then, there is still need for more research that focuses on suprasegmentals within the context of L2 development. Research studies that have focused on segmental features so far have covered topics such as perception and production of phonemic contrasts such as /ɹ/-/l/, /t/-/d/, or /p/-/b/ (e.g., Bradlow, 2008; Flege, 1995; Flege & Port, 1981), perception and production of vowels (e.g., Flege, Bohn, & Jang, 1997; Wang & Munro, 2004), and linguistic theories such as the role of universal markedness and the influence of L1 on L2 production (e.g., Broselow, Chen, & Wang, 1998; Eckman, 1981; Eckman & Iverson, 2013; Major & Faudree, 1996). The extensive focus on segmental features and setting the native-speaker model as a developmental goal, at least until recently, also led to an interest in the effects of maturational constraints on foreign accents. A common finding put forward by studies that focused on phonological development and maturational constraints is that successful acquisition of phonology decreases with age (e.g., Guion, Flege, Lin, & Yeni-Komshian, 2000; Munro, Flege, & MacKay, 1996; Munro & Mann, 2005). Although researchers such as Flege also mentioned that phonological development as an adult is still possible, these studies partially support the critical period hypothesis (Scovel, 2000), which argues that native-like speech is not achievable if L2 acquisition starts after puberty (Birdsong, 2006). However, it needs to be mentioned that this trend is changing as more and more L2 pronunciation researchers have been echoing Abercrombie’s (1949) recommendation that a useful and realistic goal for L2 learners is intelligible speech not native-like pronunciation (e.g., Derwing & Munro, 2005; Levis, 2005; Scales, Wennerstrom, Richard, & Wu, 2006). The role of intelligibility and comprehensibility as a useful developmental goal has been consolidated with the seminal works of Derwing and Munro, which showed that accentedness does not necessarily lead to low intelligibility or comprehensibility (Derwing & Munro, 1997; Munro & Derwing, 1999). Proponents of the intelligibility principle argue that if an L2 speaker can be understood without difficulty, whether he or she sounds like a native speaker would have little importance. There is evidence that training on segmental and suprasegmentals can both facilitate more intelligible and comprehensible speech. In fact, some studies argued that suprasegmentals may have a bigger role in promoting intelligibility and comprehensibility (Derwing &

Rossiter, 2002; Field, 2005; Hahn, 2004; Munro & Derwing, 1999). Yet, despite the potential benefits of suprasegmentals on L2 speech, a relatively small group of studies have explored questions related to the acquisition of suprasegmentals by L2 learners which has been highlighted in the past by various scholars (e.g., Piske, Mackay, & Flege, 2001; Trofimovich & Baker, 2006). The situation has not improved substantially since then. What is more, there are even fewer studies that have explored development of L2 phonology from an interactionist perspective including the pilot study for this dissertation (Parlak & Ziegler, 2017; Saito, 2015a, 2015b; Saito & Lyster, 2012a, 2012b). Therefore, more studies are needed to explore questions that focus on the development of suprasegmentals from an SLA perspective, particularly following the interactionist approach.

3.1. Theory driven L2 phonology studies

A considerable number of L2 phonology studies conducted during the early years of the SLA field were theory-driven. The influence of behaviorism and contrastive analysis on the SLA field, which at the time placed a greater emphasis on syntax and morphology, was also observed in research that focused on L2 phonology. Lado's (1957) seminal book introduced the contrastive analysis hypothesis (CAH) in the following words:

> We assume that the student who comes in contact with a foreign language will find some features of it quite easy and others extremely difficult. Those elements that are similar to his native language will be simple for him, and those elements that are different will be difficult. (Lado, 1957, p. 2)

The application of the CAH to predict acquisition patterns in L2 phonology followed the assumption that L2 sounds that are different from L1 sounds are more difficult to acquire. For example, the voiceless bilabial stop /p/ is not a part of the phonological inventory of Arabic but the English phonological inventory includes this particular consonant. Therefore, according to the CAH, speakers of Arabic would have difficulty producing English words with the sound /p/. For example, they would pronounce *public* as /bʌb.lɪk/. Lado (1957) also argued that separating two allophones in the learner's L1 into two distinct phonemes in the L2 would be the greatest challenge. Much of the earlier pronunciation research motivated by the CAH focused on a point-by-point comparison between various L1 and L2 phonological inventories in order to predict potential learner errors (e.g., Hammerly, 1982). However, the popularity of the CAH did not last long. Many researchers argued that the CAH lacked predictive power, and therefore, questioned its usefulness (e.g., Oller & Ziahosseiny; 1970; Wardhaugh, 1970; Whitman & Jackson, 1972). One of the shortcomings of the CAH was that it was based on contrasts that were binary. In other words, it would predict that a particular phoneme in the target language is either easy or difficult to acquire without considering the possibility that there could be different degrees of difficulty due to intermediating factors. In an attempt to address this problem and revise the CAH, Eckman (1977) proposed the markedness differential hypothesis (MDH), which argued that phonological features that are typologically marked will be more difficult for L2 learners to acquire. Eckman argued that a universal typological rule that applies to all languages needs to be incorporated into the CAH to increase its predictive power as well as allow interpretations based on degrees of difficulty. The concept of markedness as a universal typology goes back to Greenberg's (1966) discussion on universals in phonology. His analysis of nasal and oral vowels showed that the frequency of nasal vowels in a language is always lower than the frequency of oral vowels in that language. Therefore, he explained that oral vowels would be unmarked, mainly basic, and nasals vowels would be the marked form. It is important to note that markedness is a unidirectional and systematic concept. That is to say, if a language has the voiceless bilabial stop /p/, the marked form, then it will have the voiced bilabial stop /b/, the unmarked form. However, when a language has /b/, it may not have /p/, as it is the case for Arabic. In short, marked forms can be used to predict the presence of unmarked forms but not the other way around. According to the MDH, there are three main levels of difficulty for a language learner:

(a) Those areas of the target language which differ from the native language and are more marked than the native language will be difficult.

(b) The relative degree of difficulty of the areas of the target language which are more marked than the native language will correspond to the relative degree of markedness;

(c) Those areas of the target language which are different from the native language, but are not more marked than the native language will not be difficult. (Eckman, 1977, p. 321)

In a subsequent study, Eckman (1991) introduced the structural conformity hypothesis (SCH), which is an extension of the MDH that makes predictions on not only native and target languages but also interlanguages. Although it is difficult to reject the possibility of a transfer effect from the learner's L1, the CAH, the MDH, and the SCH all had a major shortcoming as a structural approach: the learner as an individual was not a part of these hypotheses.

In order to address the role of the learner in within the context of L1 influence, Flege (1995, 2003) introduced the speech learning model (SLM). There are three important characteristics of the SLM. The first is that it is a model that puts perception before production. According to the SLM, L2 learners need to create distinct perceptual categories for target language sounds to be able to produce them. In other words, accurate perception is necessary for accurate production. In Flege's (1995) words, "without accurate perceptual 'targets' to guide the sensorimotor learning of L2 sounds, production of the L2 sounds will be inaccurate" (p. 238). That said, the model still treats perception and production as separate mechanisms and acknowledges the possibility that some production errors will occur regardless of accurate perception. Secondly, the model claims that L2 phonemes that are perceptually similar to the L1 phonemes, whether it is a single L2 phonemes or two L2 phonemes that are allophones in the L1, will be difficult for the learner learn. On the other hand, the SLM predicts that it will be easier for learners to create a new perceptual category for L2 phonemes that are perceptually different from L1 phonemes. The support for this claim comes from earlier work on perception of English vowels by native speakers of German (Bohn & Flege, 1990) and native speakers of Spanish (Flege & Bohn, 1989), and perception of French vowels by native speakers of English (Flege, 1987). Finally, the SLM claims that "the phonetic systems used in the production and perception of vowels and consonants *remain adaptive over a life span"* (emphasis in original, Flege, 1995, p. 233), which means that it is possible for adult L2 learners to acquire target sounds successfully. This claim challenges the critical age hypothesis (Scovel, 2000) which argues for an effect of fossilization that occurs around the age of puberty preventing the acquisition of native-like speech at a later age. It is important to note that Flege does not deny the influence of maturational constraints on L2 acquisition. One of the hypotheses of the SLM is that "the likelihood of phonetic differences between L1 and L2 sounds, and between L2 sounds that are non-contrastive in the L1, being discerned decreases as [age of learning] increases" (Flege, 1995, p. 239). Subsequent work by Flege and his associates also confirms the role of maturational constraint in the SLM (e.g., Guion et al., 2000; Munro et al., 1996). Yet, according to the SLM, it may still be possible to overcome maturational constraints as an adult.

Some of the arguments put forward by Flege were shared by other scholars who studied L2 speech perception around the same time. Best, McRoberts, and Sithole's (1988) proposed a four-category model for *phonemic perception* of non-native sounds. They hypothesized that whenever possible, target language sounds would be assimilated into an L1 category, and it would be difficult for non-native listeners to distinguish between phonemic contrasts in a target language if those phonemes have been assimilated into L1 categories. On the other hand, they argued that phonemes that are dissimilar to any of the L1 categories would not be assimilated and that non-native listeners would be able to distinguish between these phonemes. Their study showed that L1 English speakers were able to perceive phonemic contrasts between different click phonemes in Zulu, as English does not have click phonemes. Later, Best (1995) built onto this initial work and introduced the perceptual assimilation model (PAM) which provided a framework for categorizing non-native sounds based on the listener's L1 sounds. The PAM is a model designed to predict assimilation of new phonemes during the early stages of L2 acquisition. Later, Best and Tyler (2007) expanded the model to PAM-L2 to predict different patterns of assimilation at advanced stages of L2 acquisition. They also highlighted some of the similarities and differences between the PAM-L2 and the SLM, one important difference being that “PAM-L2 addresses equivalence not only at the phonetic level addressed by the SLM, but also at the phonological level” (Best & Tyler, 2007, p. 27).

In summary, earlier hypotheses such as the CAH and the MDH treat sounds in a language as a set of objects to be analyzed ignoring the learner-related factors. These earlier approaches to L2 speech development are not useful from an SLA perspective as empirical studies have shown that “language learning is variable in its outcome” due to learner-related factors (Van Patten & Williams, 2015, p. 10). The PAM and the SLM, on the other hand, are perception-based and process-oriented approaches that at least acknowledge the learner as a factor. Therefore, as Derwing and Munro (2015) have also pointed out, the implications from these two recent models have the potential to be more useful for language learners and teachers. Yet, it should also be noted that both SLM and PAM mainly focus on segmental features, that is consonants and vowels, and they do not provide any implications regarding the acquisition of suprasegmental features such as stress and intonation.

3.2. Pronunciation instruction studies

In addition to research that pursued investigation of theory-driven hypotheses, L2 phonology studies have also explored the impact of various factors such as perceptual training or explicit pronunciation instruction on L2 learners’ phonological development. To begin with studies that focused on the relationship between perceptual training and improvement in phonological output, the departure point for the majority of these studies was the claim that accurate speech production would not occur without accurate speech perception (Best, 1995; Bohn & Flege, 1992; Flege, 1987, 1995). This line of research studies postulated that an improvement in the perception of a particular target phoneme would lead to an improvement in the production of that phoneme. A number of studies have in fact provided evidence which showed that perceptual training can have an impact on production accuracy. These studies mainly focused on the perception and production of contrast between various consonants and vowels such as the English /r/-/l/ by native speakers of Japanese (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997), Spanish /p/-/b/ by native speakers of English (Zampini, 1998), French /ə/-/ø/ by native speakers of English (Brosseau-Lapré, Rvachew, Clayards, & Dickson, 2013), French nasal vowels /ɔ̃, ɑ̃, ɛ̃/ by native speakers of English (Inceoglu, 2016), American English mid and low vowels /æ, ɑ, ʌ, ɔ, ɜ/ by native speakers of Japanese (Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005), and Canadian English vowels /i, ɪ, e, ɛ, æ, ɒ, ʌ, o, ʊ, u/ by native speakers of Mandarin (Thomson, 2011). Although less common, there were also studies that focused on the perception and the production of phonemic contrasts based on tones such as Mandarin tone contrasts by native speakers of English (e.g., Wang, Jongman, & Sereno, 2003). There were two main findings that were shared by the majority of these studies. The first one was that perceptual training helped learners to perceive the contrasts between the target phonemes. Secondly, perceptual training also led to an improvement in learners' production of the target phonemes. These findings were supported by other studies that mainly investigated the relationship between perception and production without an element of training yet found a link between the two (e.g., Bion, Escudero, Rauber, & Baptista, 2006; Flege, 1993; Flege et al., 1997). Although the majority of the studies mentioned above argued for a connection between perception and production, there were also studies that did not find a relationship between the two skills and argued that they develop separately (e.g., Zampini, 1998). Baker and Trofimovich (2006) pointed out that the differences between studies that focus on the perception-production relationship could be due to other factors such as age of arrival.

There are also pronunciation studies that investigated improvements in learners' speech development as a result of explicit instruction. A quick glance at these studies show that there is a considerable amount of variation in methodology. For example, Couper (2003) provided pronunciation instruction that focused on a range of segmental and suprasegmentals features (e.g., phonemes, word stress, sentence stress, joining sounds) to L2 learners of English in New Zealand from a variety of L1 backgrounds. The posttest revealed that there was a decrease in the number of errors made by the students. In this particular study, it was the researcher who evaluated the errors on the pretest and the posttest. Lord (2005) conducted a study that focused on the acquisition of Spanish consonants by native English speakers. The participants were enrolled in a phonetics course and the intervention was in the form of oral practice, transcription practice, and phonetic analysis of the target sounds including participants' own production. Once again, the findings indicated that participants benefitted from the intervention; however, it is important to note that the control group in this study comprised 10 native speakers of Spanish; in other words, there was not a true control group. Lord measured development through acoustic analysis of voice onset time. Another study that focused on Spanish consonants was conducted by Kissling (2013). The participants in her study were L1 English speakers enrolled in introductory, intermediate, and advanced level Spanish courses. Similar to Lord's (2005) study, participants received phonetics instruction; however, in this case there was a true control group. Kissling used listener ratings and acoustic analysis to measure development. The results did not yield any significant development for any of the levels. Akita (2005) conducted a study with L1 Japanese learners of English that compared explicit instruction focusing on segmental features to explicit instruction focusing on suprasegmental features. The results showed that the group that received suprasegmental instruction outperformed the group that received segmental instruction as well as the control group. However, Akita (2005) mentioned that in addition to differences in instruction, there was also an element of corrective feedback as well as “explicit information regarding what is ungrammatical” (p. 18). From the details provided in the article, it is not clear how explicit instruction and corrective feedback were provided. There is also no elaboration on the link between grammar and development of segmentals or suprasegmentals. In another study that tested the effect of computer-assisted training on the development of prosody, Hardison (2004) found that L1 English learners of French not only benefitted from training but they were able to generalize the improvement to segmental features and novel sentences. The training required participants to read sentences into a microphone connected to a computer, which would then provide visual and auditory feedback by comparing prosodic patterns produced by participants to those produced by native speakers of French. Learners’ pretest and posttest productions of segmental and suprasegmental features were rated by native speakers of French using 7-point rating scales. As can be seen from these examples, there is a wide range of methodological differences among L2 pronunciation studies including what constitutes as training and how development is measured.

There are also studies that focused on intelligibility and comprehensibility (Derwing & Munro, 1997) as an outcome measure. Derwing and Rossiter (2003) conducted a 12-week study with three experimental groups; one that received pronunciation instruction focusing on segmentals, one focusing on suprasegmentals, and one with no pronunciation focus. The participants were ESL university students in Canada with various different L1 backgrounds. Participants’ pretest and posttest performance judged by six expert listeners, who were native English-speaking ESL teachers, showed that the suprasegmental group showed improvement in comprehensibility and fluency whereas the segmental and the control groups did not show any gains in these two areas. The segmental group made gains in accurate production of phonemes; however, those gains did not impact their comprehensibility and fluency scores. Based on these results, the authors recommended a stronger focus on suprasegmental features when integrating pronunciation into language instruction. The findings from this study provided further support for a previous study conducted by Derwing, Munro, and Wiebe (1998), which showed that the experimental group that received suprasegmental instruction improved their comprehensibility and fluency judged by their performance on extemporaneous narratives. In another study, Parlak (2010) investigated the effects of pronunciation instruction provided over a 6-week period on learners' intelligibility and comprehensibility. The participants were ESL learners studying at an intensive English program in the United States with a range of L1 backgrounds. The instruction focused on both segmentals and suprasegmentals with suprasegmentals getting more attention judged by the amount of time spent on the target feature (approximately 30% on segmentals and 70% suprasegmentals). The raters were 18 native speakers of English who were undergraduate students at the same university. Participants' intelligibility was judged using a transcription task and comprehensibility was judged using a 7-point likert scale. The findings showed that pronunciation instruction had a positive impact as the intervention group had higher intelligibility and comprehensibility scores on the posttest. More recently, Gordon and Darcy (2016) conducted a study that also compared two types of explicit pronunciation instruction. Their participants were divided into three groups; one that received explicit instruction on suprasegmental features, one that received explicit instruction on four vowels, and a control group.

Their participants were ESL students in intact classrooms at a U.S. University. They used acoustic analysis to measure changes in segmental features and L1 English speaking non-expert raters to measure changes in comprehensibility. After a three-week instruction period, the posttests showed that the group that received suprasegmental instruction was the only group to improve their comprehensibility ratings. Although the segmental group improved their production of vowels, as measured by the acoustic analysis, their gains did not transfer over to their comprehensibility ratings. These findings are directly parallel with findings from Derwing and Rossiter (2003). Although Parlak (2010) did not have separate intervention groups for segmental and suprasegmental instruction, the common conclusion reached by these studies is that instruction focusing on both segmentals and suprasegmentals can be beneficial; however, it would be better to place more emphasis on suprasegmental features if the goal of instruction is to improve learners' comprehensibility.

The majority of the studies mention above were included in a recent review carried out by Thomson and Derwing (2015) aimed to investigate the effectiveness of pronunciation instruction. Their review of 75 studies showed that 53% percent of the studies focused on segmentals, 23% on suprasegmentals, and 24% on both. According to these numbers, once again it is clear the there is an imbalance among pronunciation studies in terms of their focus as segmental features have been getting more attention. Despite the evidence indicating that suprasegmental features have the potential to be more useful for L2 speakers, currently there is relatively less interest in exploring research questions targeting the development of stress, intonation, and rhythm or how they impact learners' comprehensibility and intelligibility.

3.3. Interactionist L2 phonology studies

Interactionist studies that focus on L2 phonology investigate issues related to the development within the context of conversational interaction. From an interactionist perspective, learners need to make pronunciation errors to receive corrective feedback during meaning focused activities. This is different from explicit instruction studies that provide a set of consciousness raising activities in order to raise learners' awareness of their pronunciation errors and allow them to practice pronunciation of target sounds. Perhaps one advantage of explicit instruction over an interactionist approach would be the focus on explicit description of motor skills required to produce a particular phoneme (e.g., physically showing the position and the function of the tongue when producing interdental fricatives). Implicit corrective feedback moves such as recasts are not able to serve a similar function. However, it is also important to note that feedback on motor skills would be more relevant in the case of segmental features and that they are less relevant in the case of suprasegmental features. Having said that, if necessary, interactionist studies can make use of more explicit types of feedback such as meta-linguistic feedback in order to explain how the learner needs to move his/her lips or tongue to produce a particular phoneme. Currently, many questions related to the relationship between interaction and the development of L2 phonology are waiting to be answered. There is only a handful of interactionist studies that have focused on development of L2 phonology and some of those studies did not focus on the effects of corrective feedback.

For example, Bueno-Alastuey (2013) conducted an observational study that investigated interaction and language related episodes that took place during voiced-based computer-mediated interaction among three different learner dyads. The dyads were same L1 (Spanish-Spanish), different L1 (Spanish-Turkish), and non-native speaker with native speaker (Spanish-American). The analysis of the interaction between learners showed that the highest number of language related episodes occurred in the different L1 group, and the majority of those language related episodes (35%) focused on phonetic issues. Also, in all three groups recasts were the most common type of corrective feedback with different L1 group utilizing it more than the other two groups. Although this study did not focus on the impact of interaction on phonological development, it provides two pieces of critical findings that underscore the role of recasts and phonetic issues during interaction. In another study, Trofimovich et al. (2014) explored the effects of interactive alignment (Garrod & Pickering, 2009) that occurred during peer interaction on primary stress placement in L2 English. The analysis of the four collaborative tasks that were given to participants, who spoke a variety of L1s, showed that participants produced the target vocabulary with correct stress placement immediately following their interlocutor's correct model. The authors also reported that the majority of the alignments occurred in the case of different target words with the same stress pattern. When it comes to studies that focused on perception, Saito and Wu (2014) investigated the effects of form-focused instruction (FFI) and corrective feedback on L1 Cantonese speakers' perception of Mandarin tones. There were two experimental groups, a FFI group versus a FFI plus corrective feedback group, and a control group. The training session comprised four activities, three of which focused on perception and one on production. Corrective feedback was operationalized as recasts that were provided when learners made a production error. Their findings showed that the FFI group improved their perception significantly on the posttest and they were able to generalize their improved perception to untrained words. In contrast, the FFI plus corrective feedback group only marginally improved their performance on the trained words and they were not able to generalize their improvement to untrained words. Therefore, the authors concluded that the learning achieved by the FFI plus corrective feedback group could be categorized as lexical learning rather than phonological development. In another interactionist study that focused on perceptual development, Lee and Lyster (2016a) investigated the effects of FFI and corrective feedback on L1 Korean speakers' perceptual accuracy of the English /i/-/ɪ/ phonemic contrast. Participants were randomly assigned to one of the two experimental groups: the FFI group and the FFI plus corrective feedback group. When participants in the FFI plus corrective feedback group made a perception error (e.g., writing down *beat* when the stimulus was *bit*), they first received feedback in the form of a repetition (e.g., *beat?*). If no self-repair occurred following the first feedback move, participants received explicit correction (e.g., *I said "bit" but you wrote "beat."*). The findings from this study showed that FFI plus corrective feedback group outperformed the FFI-only group on both the immediate and delayed posttests showing that corrective feedback facilitated improved perception of the target phonemic contrasts. Based on the findings, it can be concluded that there is still a need for more studies to understand the relationship between corrective feedback and perception of L2 sounds. It should also be noted that the study conducted by Saito and Wu (2014) targeted the perception of a suprasegmental feature, whereas Lee and Lyster (2016a) investigated the perception of segmental features. It is possible that corrective feedback affects the perception of segmentals and suprasegmentals differently; however, this needs to be tested in future studies. Despite the fact that the findings are currently inconclusive, the studies discussed so far in this section provide useful insights into how interaction and corrective feedback may affect the development of L2 phonology. The study by Bueno-Alastuey (2013) showed that negotiation for phonological issues was highly common during interaction between non-native speakers, and regardless of the dyad structure, recasts were the most common type of corrective feedback. On the other hand, Trofimovich et al. (2014) found that interactive alignment that occurs during meaning focused tasks has the potential to facilitate correct placement of primary stress placement. Finally, the study by Lee and Lyster (2016a) indicate that corrective feedback has the potential help learners distinguish phonemic contrasts in their L2, whereas Saito and Wu (2014) showed that this effect was not strong in the case of perception of Mandarin tones. Although none of these studies focused on the impact of corrective feedback and the development of productive L2 phonology, they provide highly valuable findings related to other aspect of phonological development that could guide future studies. These studies are important due to the fact they adopted an interactionist approach when addressing questions related to L2 phonology. When we turn to studies focused on the effects of interaction on production, there are a few pioneering studies that provide encouraging findings.

There are only a handful of studies that specifically focused on productive L2 phonological development using an interactionist approach. A series of studies were conducted by Saito and Lyster in an attempt to explore this understudied area. Saito and Lyster (2012a) investigated the effects of FFI and corrective feedback, operationalized as single-word recasts, on the development of the approximant /ɹ/ by L1 Japanese learners of English. The participants were assigned to three groups: FFI only, FFI plus recast, and control. The group of participants that received FFI plus recasts showed significant gains from the pretest to posttest in their production of /ɹ/, which was measured acoustically against a baseline based on native listener perceptions of /ɹ/ as well as a baseline created by using native speaker productions. On the other hand, the FFI group and the control group did not show any improvement. In another paper from the same study that reported the results on the development of /æ/, Saito and Lyster (2012b) showed that FFI coupled with pronunciation-focused recasts led to more native-like production of high-front vowel /æ/. Later, Saito (2013a) conducted another study that focused not only on the production /ɹ/ but also the perception of the /ɹ/-/l/ contrast. The participants in this study were randomly distributed into three groups: FFI only, FFI plus recast, and control. Participants' perception of the /ɹ/-/l/ contrast was measured by a forced-choice task and their production of /ɹ/ was measured by a control production task as well as a spontaneous production task. The results showed that although both experimental groups improved their perception and production of /ɹ/, it was only the FFI plus recast group that was able to generalize the improvement to untrained target words. However, in a follow-up study with a similar design, Saito (2015a) found that although the FFI only group both improved their production and perception of /ɹ/, the FFI plus recast group only showed gains in production scores but not perception scores. Saito interpreted this unexpected finding by referring to developmental readiness and argued that it could be due to participants' unreadiness that they were not able to fully benefit from the feedback. Saito conducted other studies with similar FFI designs to tease out the relationship between explicit instruction and the development of /ɹ/ (Saito, 2013b) as well as the amount of recasts received and output opportunities (Saito, 2015b). These studies showed that the combination of explicit instruction and FFI that incorporates recasts as corrective feedback led to the highest gains, and that higher amounts of recasts lead to higher instances of modified output, which affects phonological development positively. In summary, Saito's investigation of FFI, recasts, and the development of /ɹ/ through a series of studies provides one of the most detailed accounts of how interaction and corrective feedback can impact phonological development in L2.

3.4. The case for recasts and lexical stress

There is a direct relationship between learners' speech production and their intelligibility. In fact, communication breakdowns due to phonological issues, particularly mispronunciation, are known to be common in L2 oral communication (Williams, 1999). It is not surprising that pronunciation errors have high communicative value and that they may lead to communication breakdowns. Studies have also shown that negotiation for meaning that results from phonological problems is common in oral communication. For example, in her study that explored the effects of synchronous voice-based computer mediated communication on interaction, Bueno-Alastuey (2013) found that of all the language related episodes (LRE) that participants engaged in, the highest number belonged to LREs with a focus on negotiation for phonological features. Other research studies have also shown that negotiation for phonological features through interaction is common (e.g., Bitchener, 2004; Mackey et al., 2000; Mackey et al., 2007). Equally important, when L2 learners receive clarification requests or corrective feedback targeting a phonological feature, they are able to understand the intent of feedback (Carpenter et al., 2006; Lyster & Saito, 2010; Mackey et al., 2000). Some of these studies showed that learners are more likely to notice recasts that are directed at phonological and lexical errors than morphological errors (Carpenter et al., 2006; Mackey et al. 2000). Carpenter et al. (2006) interpreted this relationship between pronunciation errors and recasts by highlighting the possibility that "phonological and lexical errors are higher in communicative value and more likely to cause communication breakdowns than morphosyntactic ones, which would again increase the saliency of recasts [on] these errors" (p. 228). The increased saliency of feedback targeting phonological features and the increased chances of noticing could also be a reason for successful uptake and repair that occur following corrective feedback that focuses on pronunciation (Ellis, Basturkmen, & Loewen, 2001; Lyster, 1998b; Sheen, 2006). Taken altogether, previous studies have shown that L2 learners' notice phonological feedback, and therefore, they are likely to pay attention to the target phonological feature and benefit from feedback.

That said, as it has been discussed in Chapter 2.5, there are various feedback moves to choose from when providing feedback on learner errors and they may not be equally effective on every target language feature. When providing feedback on phonological errors, recasts would be an appropriate choice compared to other types of feedback such as output-prompting feedback or metalinguistic feedback for various theoretical and practical reasons. One of the reasons is that simultaneous provision of positive and negative evidence would be more necessary in the case of phonological errors. At this point, it may be necessary to clarify what is meant by phonological errors. The type of phonological errors discussed herein are non-target-like realizations that occur due to a mismatch between the phonological representation of the word in the learner's interlanguage and its actual realization in the target language. These non-target-like realizations may occur due to various reasons such as perceptual errors or cross-linguistic influence. In other words, phonological errors refer to consistent non-target-like production, which excludes occasional performance deficiencies and mistakes such as slip of a tongue. The reason why positive evidence is essential when addressing phonological errors is that learners may not be able to self-repair a pronunciation error upon receiving an output-prompting feedback move that pushes them to produce the target-like form. What is more, when learners are unable to fix the error on their own but are still pushed to provide the correct realization without any positive evidence to act as a model, it is unlikely that they take advantage of avoidance as a conversational strategy (e.g., Hulstijn & Marchena, 1989; Schachter, 1974) to tackle the communication breakdown. Although avoidance can be utilized as a strategy when commutation breakdowns occur due to lexical or syntactic issues, it is difficult for learners to avoid the production of certain phonemes or phonological features, as languages have a limited set of phonemes and phonotactic rules in their repertoire that speakers of that language will have to use one way or another during oral communication. Therefore, unable to produce or avoid the target-like form when faced with an output-prompting feedback move, learners may feel irritated, frustrated, or even overloaded as a result of repeated failed attempts at self-repair, unless the output-prompting feedback is provided as part of a form-focused instruction environment which provides explicit information about the target feature. In fact, recasts may be more advantageous over prompts even in situations when both types of feedback are provided along with form-focused instruction. In a study that sought to investigate the development of /ɹ/ by L1 Korean leaners of English, Gooch, Saito, and Lyster (2016) compared recast and prompts provided as part of form-focused instruction. Although both types of feedback led to phonological development, learners who received prompts produced /ɹ/ in a way which showed that the production was affected by interlanguage influence, whereas learners who received recasts were able to produce /ɹ/ in a more target-like way. The authors attributed the differences in developmental patterns between the two groups to the function of recasts as a model, in other words, to the positive evidence provided by recasts. The prompt group was left to their own device when figuring out how to correct their pronunciation, while the recast group was provided with a model that learners could repeat and learn from. As a result, their production of the target form was more accurate. When it comes to metalinguistic feedback, although it could arguably be less frustrating for learners due to the fact that it provides an explanation for what went wrong and how the target language feature is actually produced, it generally interrupts the natural flow of communication, and therefore, may take the focus away from meaning. In short, unlike output-prompting feedback moves or metalinguistic feedback, recasts can be more useful when targeting phonological errors because they provide a model and are less likely to disrupt the natural flow of communication.

Recasts are potentially less frustrating for learners because instead of pushing them to produce a phonological feature that they are unable to produce, recasts provide positive evidence by modeling the correct pronunciation as well as negative evidence by signaling the erroneous production through additional emphatic stress or use of intonation (Leeman, 2003; Saxton, 1997, 2005). A recent study in China by Huang and Jia (2016) that sought to investigate teacher and student perceptions of corrective feedback showed that students actually prefer recasts over prompts as recasts are less demanding and less challenging. Although cultural factors may have an impact on their results, it is possible that learners in other contexts feel the same way. As Saito (2013a) notes "phonological recasts can provide students with an explicit signal of errors (i.e., *enhanced negative evidence*) and with teacher pronunciation models (i.e., *enhanced positive evidence*)" (p. 520). Saito and Lyster (2012b) emphasize this unique advantage of recasts and suggest that "recasts not only enable students to notice that their pronunciation form was not intelligible enough, but also encourage them to practice their pronunciation while listening to teachers' model pronunciation" (p. 396). As pronunciation errors can be embarrassing and frustrating for many L2 learners, including the ones who have a good command of grammar and vocabulary (Derwing & Munro, 2009), the dual role of recasts can be particularly valuable when addressing pronunciation errors. When learners are not nervous or frustrated about their oral production, they may also have a better chance of focusing their attention on the feedback without feeling threatened by it. What is more, as recasts are minimally intrusive (Doughty, 2001; Gass & Mackey, 2006), they are less likely to disrupt the natural flow of communication. Because of these reasons, recasts are an ideal form of corrective feedback when addressing L2 learners' misplacement of lexical stress. That said, as it has been mentioned earlier in Chapter 2.6, recasts come in various forms. Therefore, it is also important to identify the characteristics of recasts that would be more useful in the case of phonological errors, in particular lexical stress, as it is the focus of the current study.

Explicit recasts, that is recasts enhanced with additional prosodic emphasis, would be an ideal choice when addressing phonological errors as research has shown that recasts that are enhanced with emphatic stress can trigger noticing, learner uptake, and as a result lead to language development (Doughty & Valera, 1998; Leeman, 2003; Loewen & Philp, 2006; Sheen, 2006). When treating lexical stress errors, it is possible to manipulate the prosodic structure of a recast and make it more explicit by placing extra emphasis on the correct syllable which could be in the form of increased pitch or intensity. This added emphasis has the potential to increase the salience of recasts and facilitate noticing. In addition, shorter recasts with minimal changes will potentially be more effective when providing feedback on lexical stress errors. The length of a recast and the number of changes is known to affect the degree of explicitness and the usefulness recasts (Egi, 2007). Philp (2003) found a direct correlation between the length of a recast and working memory load, and reported that the learners in her study were more likely to recall shorter recasts and hence benefit from them. Subsequently, other scholars have also argued that declarative, short, and reduced recasts are useful for L2 development (Nassaji, 2009; Sheen, 2006). Building on these findings, it would be easy to keep recasts short and focused when providing feedback on lexical stress errors, particularly if the target word is repeated in isolation. When presented in isolation, a single-word recast is a relatively short unit of speech and additional emphasis on the correct syllable would increase the contrast between the recast and learners' erroneous production of lexical stress. This type of immediate contrast is likely draw learners' attention to the target syllable and allow them to interpret the corrective force of recasts because they are less likely to be ambiguous, and therefore, can create an opportunity for L2 learners to successfully extract the negative evidence from recasts (Ellis & Sheen, 2006; Nicholas, Lightbown, & Spada, 2001). Considering all of the points above, isolated single-word recasts in declarative form that have only one change (e.g., primary stress placement) carrying extra emphatic stress are likely to be effective when providing feedback on lexical stress errors. In fact, these particular characteristics are similar to the characteristics of recasts that occur in the classroom. As Sheen (2006) reported, the majority of the recasts that occurred in her classroom study were "short, more likely to be declarative in mode, reduced, repeated, with a single error focus, and involve substitutions rather than deletions and additions" (pp. 386-387). Based on these observations, the type of recasts that the current study deems appropriate for lexical stress errors also have ecological validity.

Despite the frequency of recasts in naturalistic as well as classroom settings, and the mounting evidence that indicate their usefulness for L2 development, there is a paucity of studies that have investigated the relationship between recasts and the development of L2 phonology. So far studies have shown that learners notice the negative evidence and attempt for self-repair following recasts on phonological errors (Carpenter et al., 2006; Ellis et al., 2001; Kim & Han, 2007; Lyster, 1998a; Mackey et al., 2000). Furthermore, a number of interactionist studies by Saito and Lyster have explored this relatively less studied area through a series of studies that focused on the relationship between recasts and the development of segmental features (Saito & Lyster, 2012a, 2012b; Saito, 2013a, 2015a). In general, these studies indicated a positive impact of recasts on the development of vowels and consonants. However, there are not many interactionist studies that focused on corrective feedback and L2 phonological development, particularly the development of suprasegmentals. The interaction approach can be implemented when addressing a variety of questions related to L2 phonological development, particularly when investigating the relationship between corrective feedback and perception or production of phonological features. Recently, Foote and Trofimovich (2018) made a call for interactionist studies to investigate "the potential for corrective feedback to address problematic aspects of prosody, the relationship between task design and interactional focus on pronunciation, and the role of such variables as learners' attention capacity, motivation, anxiety, or language background and identity in their ability to benefit from interaction" (p. 81). Considering their call, and all the other reasons mentioned in this section that make recasts an ideal form of corrective feedback on phonological errors, it is meaningful and timely to investigate the relationship between recasts and lexical stress, an endeavor that would contribute to filling a gap that currently exists within the interactionist strand of SLA field.

CHAPTER 4: LEXICAL STRESS

4.1. Segmentals and suprasegmentals

Phonology lies at the heart of oral communication as speakers need to combine phonemes to create morphemes and words, which are then used to build longer stretches of speech that convey thoughts and feelings. That is to say, speech production starts at the phoneme level, which is the smallest unit of speech sounds. Phonemes are contrastive, and therefore, changing a single phoneme in a word can change the meaning of the word as it is the case with minimal pairs such as *pat-pit* or *right-light*. In many L2 phonology and pronunciation studies, the term *segmentals* are used when referring to phonemes. Researchers have defined segments as phonemic units that would be obtained when stretches of speech are divided into smaller pieces (Roach, 2000; Zsiga, 2013). In the context of L2 phonology research, segmentals are mainly used when referring to vowels, consonants, and allophones. Another commonly used term in L2 phonology and pronunciation studies is suprasegmentals, also known as prosodic features. When contrasts in speech extend beyond a single segment, as it would be in the case of a syllable, a word, or an entire clause, the speech phenomena that create these contrasts are called suprasegmental features (Roach, 2000; Zsiga, 2013). The most commonly known suprasegmental features, in no particular order of importance, are lexical stress, intonation, rhythm, and tone in tonal languages. Similar to L1 speakers, L2 learners need to master both segmental and suprasegmental features at an intelligible level in order to communicate effectively in the target language.

4.2. Accent and stress

Linguistic stress, which is a suprasegmental feature, is generally defined as the relative prominence of a syllable due to being more audible in comparison to the other syllables in the same word. The relative prominence of a stressed syllable also makes it perceptually salient. As it has been explained earlier in Chapter 2.4, the type of salience that is relevant in the context of lexical stress is psychophysical salience (N. Ellis, 2018). The salience of a stressed syllable is mainly the result of the extra physical effort exerted through articulatory organs to increase air pressure and/or vibration of the vocal folds while producing the stressed syllable. This extra effort makes the stressed syllable sound more prominent in comparison to the neighboring syllables in the same word. Although there is a consensus among researchers that perceptual salience and relativity are inherent qualities of stress, there is no agreed-upon definition of linguistic stress based on its physical properties. The lack of consensus and clarity is mainly due to the rather fluid nature of stress. There are a number of phonetic features such as duration, intensity, and pitch that come together in various combinations depending on the target language in order to form what is perceived as stress, and these phonetic features differ in the way they impact morphological and syntactic elements in various languages (Fox, 2000). The difficulty of providing a comprehensive definition for stress can lead to differences in the choice of terminology, which is mainly motivated by the particular theoretical framework a scholar follows. Therefore, it is important to briefly touch upon two key terms and explain how they will be used throughout the current dissertation.

Accent and *stress* are the two most commonly used terms when referring to perceptual salience in the context of phonology and phonetics. The term accent has been used in different ways by various scholors and in many cases these different uses contradict with one another (Fox, 2000). For example, Jassem and Gibbon (1980) consider “accent as a textual, concrete, observable category and stress as an abstract, possibly lexical, analytic category” (pp. 8-9), whereas Abercrombie (1991) considers stress to be a concrete, physical manifestation of accent, which is an abstract quality that is a part of the lexicon. Apart from clear contradictions, it is also possible to see that the terms stress and accent being used interchangeably, which has actually been highlighted earlier (e.g., Leemann, 2012). All of these different uses may create confusion for the reader. Therefore, it would probably be useful to briefly touch upon these different uses and then explain how these terms are used in the current study. To begin with accent, in many L2 pronunciation studies, it is used when referring to the distinctive way a person speaks their second language which is usually influenced by their first language (e.g., Derwing & Munro, 1997, 2009). However, in the fields of phonology and phonetics, the term accent refers to auditory prominence, but what that prominence refers to may differ based on the theoretical framework one follows (e.g., Abercrombie, 1991; Cutler, 1984; Jassem & Gibbon, 1980; Liberman & Prince, 1977; van der Hulst, 2012). This section will briefly touch upon how the words accent and stress are used by scholars who follow the Abercrombian framework as well as the those who follow autosegmantal metrical phonology (AM) framework. According to the former, the word *accent* refers to an abstract element of prominence that is an inherent element of the lexicon. Abercrombie (1991) presents accent as an overarching term for prominence that cannot be explained or predicted by phonetic factors, which detaches the term from physical description of speech sounds. When it comes to stress, Abercrombie defines it as one of the various possible realizations of accent that can be explained in terms of perceptual phenomena such as pitch, duration, and intensity, which allows physical description. In a reprint of his earlier definition of accent, Abercrombie (1991) notes that:

> An accented syllable may be realised as stress, with various features of pitch,

> of syllable length and segment length, of loudness, and of articulatory characteristics in various combinations. But none of these are included in the definition of accent. In other words, accent is ineffable. It plays no part in the phonological analysis of utterances; its place is in the lexicon. (p. 82-83)

Following the Abercrombian framework, van der Hulst (2014) suggests that "the terms 'accent' and 'stress' *can* be used distinctively, accent being the term for 'substance-free' lexical marks and stress for phonetic and phonological correlates of accent" (emphasis in the original, p. 6). He also explains that accent can both lie below stress in the lexicon as an abstract idea of prominence, as well as above stress in connection to the realization of stress and as part of the intonational structure of the utterance. The treatment of the term accent as an underlying concept of prominence in the lexicon leads to categorizations such as *stress-accent* and *pitch-accent* (or sometimes non-stress-accent). These two terms are in fact two of the three main categories that form the typology used by some scholars for defining prominence at the lexical level, namely, *tone*, *stress*, and *lexical pitch accent* (Fox, 2000; Jun, 2005). When used in this sense, the terms stress-accent and pitch-accent can be used to categorize languages based on how accent is realized at the lexical level in that language (see Beckman, 1986). That is to say, what is perceived as a salient syllable can be realized through the use of pitch in languages such as Japanese, making those languages a pitch-accent language, and through the use of a combination of perceptual phenomena such as duration, intensity, and pitch in many European languages including English and German, making them a stress-accent language. It is important to highlight that pitch-accent and stress-accent are both lexically contrastive. Yet, these terms are not far from criticism either. It has been argued that the use of the term *pitch-accent* at the lexical level can be problematic because a) some languages labelled as pitch-accent may not always make use of pitch for assigning prominence as the notion of pitch-accent is not consistent but “rather a ‘pick and choose’ among the properties that characterize prototypical tone vs. stress-accent systems” (Hyman, 2006, p. 226), and b) it would require a more detailed treatment of other correlates of stress, which could possibly necessitate adaptation of terms such as duration-accent (Gordon, 2014; van der Hulst, 2014). In fact, Hyman (2009) argues against using pitch-accent as a third category altogether and suggests using the remaining two categories, namely, tone and stress-accent.

To sum up, following Abercrombie’s (1991) framework, the characteristics of accent and stress outlined by van der Hulst (2003) show how accent can be considered an abstract notion of prominence with stress being one of its physical manifestations:

(1) Cues for accent in English

a. Inherent stretchable properties (duration, pitch, loudness, manner)

b. Anchoring of intonational tones

c. Lexical-phonotactic constraints

d. Post-lexical ‘phonetic’ constraints (and the processes that serve them) (p. 4)

(2) Phonetic properties of stressed syllables

a. The stressed syllable has greater duration

b. The stressed syllable is louder (greater amplitude)

c. The stressed syllable is pronounced at a higher pitch (higher fundamental frequency)

d. The segments are pronounced with greater precision (p. 2)

The combination of the four properties of stress listed under (2) forms the first cue for accent under (1), which indicates that stress is one of the possible physical manifestations of accent. The two lists also make it clear that, for scholars who follow the Abercrombian approach, accent is a more fundamental element of prominance compared to stress.

On the other hand, some scholars use the term stress when referring to lexical level prominence and the term accent when referring to utterance level prominence, although there may still be conceptual differences in the way they use the two terms (e.g., Cutler, 1984; Fox, 2000; Ortega-Llberaria & Prieto, 2007; Sluijter & Heuven, 1996a). For example, Cutler (1984) explains that both stress and accent are concepts of relative prominence, yet at two different levels. She explains that "word stress patterns are part of the lexical identity of words, not arbitrarily assigned by rule," whereas "Sentence accent […] expresses the information structure of a sentence; when a sentence is produced the speaker assigns accent according to what he considers to be the more and less important parts of what he is saying" (Cutler, 1984, p. 89). This particular usage of the term pitch-accent corresponds to what some researchers define as sentence stress (e.g., Chen, Robb, Gilbert, & Lerman, 2001) or tonic stress (Jenkins, 2004; Seidlhofer, 2004). When used in this sense, one of the functions of pitch-accent can be highlighting a particular word when introducing new information or for purposes of lexical contrast (van der Hulst, 2003; Zsiga, 2013). In English, the pitch-accent normally falls on the last word of an utterance. However, when the speaker wants to convey a particular emphasis or contrast, the pitch-accent can be placed on an earlier word in the utterance:

(1) emphatic stress

My friend's cat is <u>so</u> cute.

(2) non-contrastive pitch-accent

I am planning to invite Roger to the party.

(3) contrastive pitch-accent

I am planning to invite Roger to the party.

In the first example, the speaker uses pitch-accent for an emphatic purpose to show how much she adores her friend’s cat. In the second example, the speaker simply shares her plan with the interlocutor without a particular emphasis. This response could be provided following the question *What are your plans?* However, in the third example, the speaker emphasizes the word *Roger,* which could mean “not you, her, or someone else.” This response could be given to the question “Are you planning to invite Jane to the party?”

Last but not least, it is important to touch upon how the terms stress and accent are used within the autosegmental-metrical (AM) framework (Pierrehumbert, 1980; Ladd, 2008) used in intonation research. Scholars who follow the AM framework investigate intonational patterns by focusing on the connection between the autosegmental tier that represents the melodic level and the metrical tier that represents stress and phrasing (Arvaniti, in press). Within the AM framework, term pitch-accent refers to pitch movements that usually co-occur with stressed syllables, which are referred to as metrical heads. In other words, pitch-accents in English can be a cue for stress. This is different from lexical pitch-accent that was mentioned earlier, as within the AM framework the term *pitch-accent* does not have a lexically contrastive function that changes meaning of a word, but it is rather a post-lexical intonational element that enhances the prominence of syllables in accordance with the metrical structure. In addition to some of the examples provided in the previous paragraph to show how pitch-accent can be used for purposes of assigning focus to highlight a particular word, AM scholars have argued that pitch-accent may sometimes fall on words due to the metrical requirements rather than contrastive emphasis or focus on new information (German, Pierrehumbert, & Kaufmann, 2006). As pitch-accents form the melody of an utterance, there can be more than one pitch-accent in an utterance. The most important pitch-accent in a phrase is called the *nuclear accent* (Arvaniti, in press; Ladd, 2008).

It is beyond the scope of this dissertation to delve further into the Abercrombian or the AM approach to accent. However, the brief overview provided in this section shows the ways in which terms stress and accent may be used depending on the theoretical framework. Understanding how these two terms are used can be helpful for conceptualizing different levels of phonological prominence. As the current study investigates prominence at the lexical level and aims to measure L2 pronunciation development acoustically, it is main focus is what some scholars call *lexical stress-accent* which refers to the emphasis placed on a particular syllable in a word which makes it perceptually more salient than the immediately surrounding syllables. For the sake of simplicity, the term lexical stress will be used instead of stress-accent. Now we turn to lexical stress, which is the primary focus of this study.

4.3. Lexical stress in English and Arabic

As a suprasegmental feature, lexical stress is contained within a syllabic unit which is a fundamental phonological structure based on hierarchical models of phonology (James, 1986). Lexical stress can be examined from two perspectives, namely, production and perception (Roach, 2000). In perceptual terms, lexical stress is defined as the emphasis given to a syllable in a multisyllabic word which causes the stressed syllable to be perceived as more prominent by the listener (Archibald, 1993; Major, 2001). It is important to highlight that the perceived prominence will be relative to the other syllables in the word (Katamba, 1996). In fact, as Ladefoged (2006) mentions "all the suprasegmental features are characterized by the fact that they must be described in relation to other items in the same utterance" (p. 24). In the case of two-syllable words, perceived prominence is a dichotomous concept; therefore, two-syllable words have one stressed and one unstressed syllable. For example, the words *WAter* and *SOlid* have stress on the first syllable, whereas *oCCUR* and *deNY* have stress on the second syllable (capitalizations indicate primary stress placement). On the other hand, words that have three or more syllables usually carry multiple levels of stress. Generally speaking, a four-level categorization is used in order to differentiate among different types of stress in multisyllabic words. These four categories are primary stress, secondary stress, tertiary stress (unstressed but not reduced), and reduced (Katamba, 1996; Zsiga, 2013). Regardless of the number of syllables in a word, there can be only one syllable that carries primary stress, or in other words, the prosodic peak. This aspect of stress is called culminativity (Heinz, 2014; Hyman, 1977). For example, in the word *imPORtant*, the first syllable carries the secondary stress, the second syllable carries the primary stress, and the third syllable is usually reduced (the vowel is pronounced as a schwa). In the word *eLIminate*, the first syllable carries the tertiary stress, the second syllable carries the primary stress, the third syllable may be reduced or have tertiary stress depending on the speaker's dialect, and the fourth syllable carries the secondary stress. Identifying the primary stress in a clearly spoken word can be a straight-forward task for even untrained native speakers. However, deciding which of the remaining syllables carry the secondary and tertiary stress may not be an easy task, as the magnitude of the contrastive differences between non-primary-stress syllables is usually not as salient as the difference between a primary-stress syllable and any other syllable.

Lexical stress can be fixed or free-flowing depending on the language. For example, in Czech and Hungarian primary stress is almost on the first syllable (Fudge, 2016), whereas in Polish it is on the penultimate syllable (Ladefoged & Disner, 2012). In languages such as Spanish (Flege & Bohn, 1989) and Russian (Chrabaszcz, Winn, Lin, & Idsardi, 2014), the location of lexical stress is unpredictable; therefore, these languages have *free stress,* or the stress in these languages is *lexical* meaning that the only way to remember the location of stress would be through memorization. In many cases, stress in English can also be lexical. For example, there is no clear rule for why the words *COmedy* and *DEsert* are stressed on the first syllable but *coMMIttee* and *deSSERT* are stressed on the second syllable. On the other hand, stress in English can also be *paradigmatic*, which means that stress assignment depends on morphological information such as whether the word is a noun or a verb (Zsiga, 2013). For example, there is a predictable stress pattern in noun-verb homographs such as "*import, insult, object, refund, segment*." There is no segmental difference in any these noun-verb pairs and their pronunciation depends on whether they function as a noun or a verb. When these words are used as a noun, the stress will be on the first syllable and when they are used as a verb, the stress will be on the second syllable. These examples show that lexical stress in English can also have a contrastive function.

Describing the production of lexical stress, in other words its phonetic properties, is a more complex endeavor than describing it in perceptual terms. That is mainly due to the fact that stress in English does not have a single acoustic correlate. Lexical stress is acoustically multidimensional and has a number of acoustic correlates that correspond to what is perceived as a prominent syllable. There is a general consensus that duration, pitch, and intensity are the main correlates of lexical stress (Beckman, 1986; Fry, 1955, 1958; Liebermann, 1960). Pitch and intensity are perceptual terms corresponding to fundamental frequency (f_0) and signal amplitude measured in decibels (dB) respectively. Although there is little disagreement on the contribution that each of the three acoustic correlates of stress makes to perceptual prominence, research so far has not been able to explain in what combinations and proportions these correlates come together to form what is perceived as a prominent syllable. Also, the predictive value of the acoustic correlates of stress can differ from one language to another. Earlier works by Bolinger (1958) and Fry (1955, 1958) identified pitch, duration, and intensity as the phonetic cues for primary stress in English. These studies argued that pitch is the main predictor of lexical stress, followed by duration and intensity, with findings on intensity being mixed. Similarly, more recent studies suggest that overall intensity may not be a good correlate of stress as there may be a possible confounding effect created by pitch-accent (Sluijter & van Heuven, 1996a). When it comes to earlier studies that identified pitch as the main correlate of stress, subsequently it has been argued that in those studies there was a misinterpretation of the function of pitch as the words were produced in focus position. This means that they attracted the intonational pitch-accent, which led to a confounding effect (Beckman, 1986; van der Hulst, 2014). On the other hand, more recent studies have argued that duration is a more reliable predictor of lexical stress than pitch is (Gordon & Roettger, 2017; Kochanski, Grabe, Coleman, & Rosener, 2005; Ortega-Llebaria & Prieto, 2010). In addition to the studies that focused on the three main correlates of stress separately, recent studies argued that alternative acoustic features can be a predictor of lexical stress such as spectral balance combined with duration (e.g., Sluijter & van Heuven, 1996a, 1996b) or vowel quality combined with duration (e.g., Zhang & Francis, 2010).

While there are contradictory findings on the role of pitch and intensity, there is almost a consensus among researchers in regards to the role of duration as a reliable correlate of stress. In their survey of 110 (sub-)studies that focused on word stress, Gordon and Roettger (2017) found that the majority of the studies, to be specific 100 out of 110, investigated the role of duration as a correlate of stress and 85 of those 100 studies concluded that duration is a predictor of stress. There is also evidence that duration is a cross-linguistic correlate of stress (e.g., Arvaniti, 2000, for Greek; de Jong & Zawaydeh, 2002, for Arabic; Dogil & Williams, 1999, for German; Ortega-Llebaria, 2006, for Spanish; Sluijter, van Heuven, & Pacilly, 1997, for Dutch; Turk & Sawusch, 1996, for English). Yet, considering the complex nature of production and perception of lexical stress, the general consensus among researchers is that lexical stress needs to be studied from a multidimensional perspective as a single correlate may not be a good predictor on its own (Cutler, 2005; Lehiste & Peterson, 1959; Lieberman, 1960). Therefore, it is crucial to focus on duration, pitch, and intensity when studying lexical stress in English, whether it is produced by native speakers or non-native speakers.

When it comes to Arabic, it is also categorized as a stress language. However, unlike English, lexical stress in Arabic is not free-flowing; it is predictable and determined based on syllable weight in all Arabic dialects (Hellmuth, 2013). In general, a super-heavy ultimate syllable (CVCC or CVVC) or a heavy penultimate syllable (CVC or CVV) will attract stress in all dialects of Arabic; however, two particular exceptions to this rule are observed in San'ani Arabic (the dialect spoken in Yemen), when stress assignment falls on a heavy penultimate or antepenultimate syllable with a CVV pattern, or when it falls on a syllable followed by a geminate consonant (Watson, 2007). Despite the occasional phonological differences in the location of stress depending on the dialect, there is evidence that the realization of stress is almost the same and mainly rely on duration, pitch, and intensity (Almbark, Bouchhioua, & Hellmuth, 2014). As a result, word-level prosody in Arabic shares some similarities with English in that duration, pitch, and intensity have been observed to interact with lexical stress (de Jong & Zawaydeh, 1999; Zuraiq & Sereno, 2007). However, Arabic differs from English in that lengthening due to consonant voicing (de Jong & Zawaydeh, 2002; de Jong, 2004; Flege & Port, 1981) or vowel reduction (Zuraiq & Sereno, 2007) are not a phonological feature. Studies that compared lexical stress placement in English by L1 Arabic and L1 English speakers indicated that both of groups use duration, pitch, and intensity; however, L1 Arabic speakers may produce higher pitch values compared to native speakers and they tend not to reduce unstressed vowels (Almbark et al., 2014; Zuraiq & Sereno, 2007). Some studies also found a heavy reliance on duration as a result of cross-linguistic influence (e.g., Munro, 1993). In short, L1 Arabic speakers may have some advantages when learning lexical stress placement in English due to the similarities between the two languages. At the same time, not reducing unstressed vowels may create challenges.

4.4. Importance of lexical stress for L2 learners

Communication breakdowns due to intelligibility issues that are triggered by various types of pronunciation errors are common in L2 oral communication (Williams, 1999). In many cases, pronunciation problems have a negative impact on L2 speakers' intelligibility, and as a result, they may hinder successful communication and lead to communication breakdowns (Hahn, 2004; Zielinski, 2008). Pronunciation problems in L2 English can occur due to phonetic issues such as voicing (e.g., *push* versus *bush*), manner or place of articulation (e.g., *right* versus *light, think* versus *sink*), or they can occur due to misplacement of stress (e.g., *OBject* versus *obJECT*). Among these various possible pronunciation issues in L2 English, it has been suggested that lexical stress carries a bigger weight than phonetic issues in terms of perceptual salience particularly from the perspective of native speakers (Bond, 1999, 2005). In English, lexical stress is contrastive and has a direct impact on meaning or intelligibility. For example, lexical stress in English can determine the grammatical function of words such as *object, contract,* or *subject* (Fry, 1955, 1958). Misplacement of primary stress can also lead to production of a totally different word, as it could turn the word *committee* into *comedy,* particularly for L2 learners who are trying to adopt the American pronunciation by producing *committee* with a flap consonant at the onset of the third syllable. It could also render a word totally unintelligible. If the word *pyjamas* is produced with primary stress on the first syllable with a full vowel, which would possibly be pronounced as /ɪ/ (e.g., /ˈpɪ-jä-məz/, the production would be almost completely unintelligible to native and non-native listeners. Bansal showed that misplaced lexical stress by Indian speakers cause the listeners to perceive *atmosphere* as *must fear, character* as *director,* and *written* as *retain* (as cited in Cutler, 1984, p.79). Based on these examples, it is clear that lexical stress is highly important in terms of its communicative value, because when misplaced, it can drastically reduce the intelligibility of a word and in some cases change the part of speech or the meaning of a word. As a natural consequence of these issues, lexical stress can have a negative impact on intelligibility and comprehensibility of L2 speech (Kang, 2010; Isaacs & Trofimovich, 2012). Studies that focused on the relationship between prosody and intelligibility and/or comprehensibility provide further support for these points.

Being one of the earliest studies that focused on the relationship between suprasegmentals and pronunciation scores, Anderson-Hsieh, Johnson, and Koehler (1992) found a strong correlation between suprasegmental scores, which included judgment of stress, and impressionistic judgments of pronunciation by experienced raters. The study design did not allow strong conclusions to be made; nevertheless, it indicated a positive relationship between higher prosodic scores and pronunciation scores. A small-scale study by Gallego (1990) with three non-native English-speaking teaching assistants showed that 35.8% of the pronunciation problems that led to lower intelligibility and communication breakdowns were related to stress errors. The three teaching assistants in the study were L1 speakers of Korean, Italian, and Hindi, and the pronunciation problems were rated by ESL specialists. In a more recent study that investigated the impact of various linguistic influences (e.g., type frequency, vowel reduction ratio, mean length of run, grammatical accuracy, pitch contour, inter alia) on experienced raters' comprehensibility judgments, Isaacs and Trofimovich (2012) found that word stress had the largest effect size among all other variables and was the only distinguishing factor among low, intermediate, and high levels of comprehensibility. In two related studies based on the same speaker data, Trofimovich and Isaacs (2012) and then Saito, Trofimovich, and Isaacs (2016) found that word stress determined listeners' judgments of accentedness and comprehensibility. These two studies mentioned that word stress had a greater impact on accentedness judgments with the impact of word stress on comprehensibility being at a considerably high degree among other variables such as fluency, lexis, and grammar. The impact of suprasegmentals and lexical stress on L2 learners' comprehensibility has also been emphasized by studies that focused on pronunciation instruction. In a study that compared pronunciation instruction with a focus on segmental elements to instruction with a focus on suprasegmental elements, Derwing et al. (1998) found that learners who received suprasegmental instruction, which addressed word stress issues as well, showed significant gains in comprehensibility and fluency ratings judged by learners' extemporaneous production. Similarly, a more recent study by Gordon and Darcy (2016), investigated the effects of different types of pronunciation instruction by using intact classrooms. One class received explicit suprasegmental instruction, which included word and sentence stress; another class received explicit segmental instruction; and the third group received pronunciation practice with no explicit instruction. The authors found that the class that received explicit pronunciation instruction with a focus on suprasegmentals showed higher comprehensibility gains on the posttest compared to the other two classes.

One possible explanation for the relationship between lexical stress and comprehensibility could be the role of stress in lexical retrieval. One of the earliest studies that argued for the role of lexical stress in organization of the mental lexicon was conducted by Fay and Cutler (1977). Their analysis of speech errors, mainly malapropisms, indicated that the wrong words shared the same grammatical category (99%) and stress pattern (98%) with the intended word. Based on the findings, the authors argued that malapropisms were highly systematic rather than being a random slip of the tongue. They also hypothesized that lexical stress patterns could play an important role in storing vocabulary in the mental lexicon as well as retrieving vocabulary from the mental lexicon during speech production. In a shadowing study aimed at studying perceptual effects of mispronunciation and lexical recovery, Bond and Small (1983) manipulated target vocabulary in recorded passages by changing voicing of consonants, frontness and backness of vowels, and lexical stress placement. Then they asked native speaker participants to repeat prose passages that contained the ill-formed target vocabulary. The results showed that participants were able to produce 58% of the mispronounced words that had altered consonants in their correct form. However, the percentage of correct restoration went down to 15% for the vowel condition and 22% for the stress condition. These studies provide support for the role of lexical stress as a key factor in the competition model of lexical access. According to the competition model, incoming speech stream activates a number of words in the mental lexicon and listeners use lexical stress cues in the speech stream to access the intended word (McQueen, Norris, & Cutler, 1994). It has been argued that lexical activation starts upon hearing the first syllable of a word as suprasegmental information is readily available in the first syllable, which is then used to search the mental lexicon for lexical access (van Donselaar, Koster, & Cutler, 2005). In line with these arguments, a number of recent studies that focused on the relationship between lexical stress and word recognition also found a positive link between the two.

For example, Cooper, Cutler, and Wales (2002) found that native English speakers make use of lexical stress cues for word recognition when processing speech input. In their study, participants were exposed to single- and two-syllable truncated forms that are segmentally identical but differed in primary stress placement such as the syllable *mu-* produced with or without stress, or *admi-* produced as *ADmi-* with primary stress on the first syllable, or as *admi-* with no primary stress to hear. The findings showed that truncated forms would activate words with matching stress patterns, which would be *MUsic* or *muSEum* in the first example and *ADmiral* or *admiRAtion* in the second, but the effect was larger for words that began with a stressed syllable. Other word recognition studies have provided similar findings (e.g., Jesse, Poellmann, & Kong, 2017; Mattys, 2000). A very recent study by Jesse et al. (2017) investigated the relationship between lexical stress and spoken word recognition by using eye-tracking technology as a methodological enhancement. Their findings provided support for earlier studies highlighting the role of lexical stress and suprasegmental information in spoken word recognition. The role of lexical stress in spoken word recognition has been supported by studies that focused on languages other than English such as Dutch (Cutler & van Donselaar, 2001; van Donselaar et al., 2005) and Spanish (Soto-Faraco, Sebastián-Gallés, & Cutler, 2001). Further evidence for the importance of lexical stress in terms of word recognition and intelligibility come from Field's (2005) seminal study which showed that a shift in lexical stress placement can render the word unintelligible not only for native listeners but also for non-native listeners. Field (2005) showed that a shift of primary stress, particularly a rightward shift, in disyllabic words impairs the intelligibility of the word for native and non-native listeners. He explained that when a word with misplaced primary stress such as *foLLOWED* occurs toward the beginning of an utterance, it may be perceived as *load* or *flowed*, which would then shape the expectations of the listener in terms of what would follow next as part of the utterance and hinder decoding of the message. His findings also concur with earlier studies that investigated stress shift in disyllabic words and argued for a negative impact of the rightward shift on intelligibility (Cutler & Clifton, 1984). In a recent paper, Lewis and Deterding (2018) discuss the issue of lexical stress in lingua franca contexts and explain how misplacement of stress can lead to misunderstandings between non-native speakers. Also, Richards (2016) found that lexical stress errors that include a vowel change hampered intelligibility and the impact was greater for non-native listeners compared to native-speakers. In sum, although not all stress shifts may hamper intelligibility equally (Levis & Wu, 2018), there is empirical support showing that listeners, whether they are native or non-native speakers, rely on lexical stress cues when processing speech input making the communicative value of lexical stress highly crucial. Based on these findings, it is highly likely that lexical stress errors play an important role in communication breakdowns as they have a direct impact on lexical access which would then hamper intelligibility and comprehensibility. This is in fact line with L2 phonology literature that has emphasized the importance of prosodic features for intelligible pronunciation (Colantoni, Steele, & Escudero, 2015; Setter & Sebina, 2018).

Lexical stress as a key phonological aspect that impacts intelligibility can present both perception and production challenges for many L2 learners. A major challenge for L2 speakers of English is the fact lexical stress in English is free flowing. Unlike languages such as French or Polish, the placement of stress in English words is not fixed and for the most part it does not follow a predictable pattern (Ladefoged, 2006). This presents a challenge for L2 speakers of English as they may not be able to correctly produce lexical stress due to its unpredictability. Misplacement of stress due to cross-linguistic influence is another challenge that L2 speakers may face. In a series of studies, Archibald (1992, 1993, 1994) investigated the perception and realization of lexical stress in English by L1 speakers of Polish, Spanish, and Hungarian. In all three studies, he found evidence for an effect of L1 transfer on perception and realization of lexical stress by the participants. In a later study with L1 speakers of Chinese and Japanese, Archibald (1997) highlighted the differences between realization of lexical stress as produced by speakers of stress languages versus non-stress languages, and he concluded that speakers of stress languages may have an advantage over speakers of tone or pitch-accent languages. On the perception side of the issue, L2 learners may employ the acoustic correlates of stress differently compared to L1 speakers when processing lexical stress information (Zhang, Nissen, & Francis, 2008). What is more, in some cases L2 learners may not be able to perceive lexical stress placement at all. For example, L1 speakers of languages such as French, which does not have contrastive stress, may struggle with distinguishing between two words that differ from each other based on stress placement (Dupoux, Peperkamp, & Sebastián-Gallés, 2001).

Nevertheless, challenges or differences in acquisition patterns should not be interpreted as lower changes of success in acquisition of prosodic features including lexical stress. Studies have shown that it is possible for adult L2 learners with different L1 backgrounds to successfully produce word stress in their L2. For example, in a study with early and late Spanish-English bilinguals, Guion, Harada, and Clark (2004) found that both groups were able to extract stress patterns from input and use the newly learned patterns when assigning stress to non-words. In another study that investigated prosodic transfer effects, Nguyễn, Ingram, and Pensalfini (2008) investigated perception and production of English stress at the word and the phrase level by beginner and advanced Vietnamese learners of English. The authors observed cross-linguistic effect in realizations by both groups; however, the advanced learners were able to manipulate duration in their productions of stress. Apart from these studies, in general there is a consensus among L2 pronunciation researchers that suprasegmental elements are learnable (Derwing et al., 1998; Derwing & Rossiter, 2003; Setter & Sebina, 2018; Trofimovich & Baker, 2006) and that pronunciation instruction facilitates phonological development in L2 (Lee, Jang, & Plonsky, 2015; Saito, 2012; Thomson & Derwing, 2015).

CHAPTER 5: THE PILOT STUDY

As the pilot study was published in a peer-reviewed journal (Parlak & Ziegler, 2017), this chapter provides only a brief summary in order to avoid redundancy and repetition. The pilot study had been conducted at the same university where the main study was conducted. The data for the pilot study was collected about one year prior to the data collection for the main study from a totally different group of L1 Arabic speaking learners. That said, both pilot study and main study learners shared the same demographics and English proficiency levels. In other words, they were recruited from the same population. The pilot study focused on the development of primary stress by comparing oral interaction in face-to-face (FTF) and synchronous computer-mediated communication (SCMC) conditions. The target vocabulary chosen for the pilot study were three- and four-syllable words with stress on the second syllable.

5.1. Method

5.1.1. Research questions

As an exploratory study, the pilot study aimed to answer the following research questions:

RQ1: What are the effects of recasts on learners' development of lexical stress?

RQ2: Compared to face-to-face contexts, how effective are recasts provided during synchronous computer-mediated video task-based interaction at promoting learners' development of lexical stress?

5.1.2. Participants

64 learners participated in the pilot study. However, six learners spoke a first language other than Arabic. These learners carried out the activities along with L1

Arabic speaking learners but their data was removed during the analysis stage. Also, the data for one learner who spoke Egyptian Arabic was removed as the learner failed to complete the tasks in accordance with the instructions. After removal of the data from these particular learners, the remaining data set comprised 57 learners who spoke Arabic as a first language. The majority of the learners were speakers for Gulf Arabic (n = 34), followed by speakers of Levantine (n = 12), Egyptian (n = 7), Sudanese (n = 2), Libyan (n = 1), and Iraqi (n = 1) dialects. 52 learners were enrolled in the Intensive English Program and 5 learners were first-year undergraduate students at the same university where the data for the main study was collected. Learners proficiency level ranged between an International English Language Testing System (IELTS) score of 5.5 and 6.5.

5.1.3. Language tasks

There were three handouts two of which were used as pre-task activities and one for the interview task. Before providing more details on the tasks and activities, it is important to clarify what is meant by the term *task* in the context of the current study, as plenty of definitions have been suggested over the years with no consensus among SLA researchers about which definition to accept. One of the earliest definitions of the term task was suggested by Long in a book chapter that provided his early thoughts on task-based language teaching:

> a piece of work undertaken for oneself or for others, freely or for some reward. Thus, examples of tasks include painting a fence, dressing a child, filling out a form, buying a pair of shoes, making an airline reservation, borrowing a library book, taking a driving test, typing a letter, weighing a patient, sorting letters, making a hotel reservation, writing a cheque, finding a street

> destination and helping someone across a road. In other words, by “task” is meant the hundred and one things people do in everyday life, at work, at play, and in between. Tasks are the things they will tell you they do if you ask them and they are not applied linguists. (Long, 1985, p.89)

A similar, non-technical approach to the term task is provided by Van den Branden (2006) who explains that “A task is an activity in which a person engages in order to attain an objective, and which necessitates the use of language (p.4). On the other hand, some SLA scholars define the term task in a more technical way, which is within the context of language learning. For example, according to Ellis:

> A task is a workplan that requires learners to process language pragmatically in order to achieve an outcome that can be evaluated in terms of whether the correct or appropriate propositional content has been conveyed. To this end, it requires them to give primary attention to meaning and to make use of their own linguistic resources, although the design of the task may predispose them to choose particular forms. A task is intended to result in language use that bears a resemblance, direct or indirect, to the way language is used in the real world. Like other language activities, a task can engage productive or receptive, and oral or written skills, and also various cognitive processes. (R. Ellis, 2003, p.4)

A shorter reformulation of Ellis’s definition is provided by Samuda and Bygate (2008) who posit that “[a] task is a holistic activity which engages language use in order to achieve some non-linguistic outcome while meeting a linguistic challenge, with the overall aim of promoting language learning, through process or product or both" (p. 69). Although there is a considerable degree of commonality among these definitions, there are also some differences depending on the theoretical underpinnings. For example, Long and Van den Branden prefer a non-technical use of the term task by referring to situations that do not necessarily occur inside the language classroom. When discussing the key characteristics of target tasks in relation to language learning and the pedagogical tasks that are derived from them, Long (2015) emphasizes the importance of establishing a direct connection to real-world activities and the needs analysis that would facilitate that connection. In other words, he emphasizes the importance of focusing on meaning and placing L2 learners' communicative needs in real-life situations at the core of task creation and task-based language teaching in general.

The other two definitions are more technical in nature in that they specifically focus on what happens or what should happen during a task within the context of a language classroom. They mainly highlight the importance of focusing on meaning through communicative use of language with the goal of reaching a non-linguistic outcome as a crucial element of tasks. The definition by R. Ellis (2003) and its reformulation by Samuda and Bygate (2008) incorporate six critical features that function as a guideline for identifying a task. The tasks in the current study meet the six criteria proposed by R. Ellis as learners in the current study had the opportunity to draw upon their personal experience and they used language creatively to reach a communicative outcome. Therefore, the tasks used for the current study can be labeled as *focused tasks* as discussed by R. Ellis (2003). On the other hand, from Long's perspective, they would probably not qualify to be labelled as a task. The reason for this is that the tasks created for the current study were not informed by needs analysis, nor had they a direct connection to what the specific group of learners who participated in the current study actually needed in real-life situations, at least not any time soon. In the context of a language classroom, they would form the elements of what Long (2015) calls *a task-supported approach* or an activity as the tasks were not driven by needs analysis, but still, provided opportunities for meaningful use of language that was driven by a communicative goal.

One last important point to mention is that the definitions provided earlier mainly discuss tasks in relation to task-based language teaching within a language classroom context. However, as Bygate, Skehan, and Swain (2001) explain, tasks can also be used for research purposes as a focused activity because they can help "generate data of interest to the researcher" (p. 6). In fact, SLA research has made use of tasks extensively to elicit language production and facilitate meaningful interaction (Van den Branden, 2006). The tasks used for the current study were developed with the same purpose. They were focused tasks that were designed to ensure consistency of target language production while allowing for learners to be creative with language when interacting with their partner or the researcher. There were two pre-task activities and three communicative tasks that were collectively built around the theme of a job interview for a teaching position. Learners first carried out a sentence-completion activity with 10 items that aimed to their ideas about language teaching and language teachers. For each item, the target word (italicized in the examples to follow) was embedded in sentence-medial position. Learners were asked to rely on their opinion and knowledge when completing the sentences. They were also informed that there were no right or wrong answers (see Example 1 on the next page).

Example (1)

- A teacher can motivate students who are *reluctant* to speak English by ________

  ____.

- A teacher needs to have *authority* in the classroom because _________________

  ____.

After the sentence-completion activity, learners carried out an information exchange task that required them to share their ideas that were written down during the sentence-completion activity. This was a communicative tasks which required learners to discuss whether they agree or disagree with their partner's ideas. After that, learners participated in a role-play interview task with a native-speaker of English who was a faculty member at the same university. The handout for the interview task contained a list of 10 interview questions that were directly related to the sentences that learners completed as part of the sentence-completion activity and the discussion that followed during the information-exchange task. Each question had one target word embedded in sentence-medial position. Learners, who assumed the role of an interviewer, were instructed to listen to the responses given by a faculty member, who pretended to be the applicant, and take notes during the interview role-play task (see Example 2).

Example (2)

- How do you motivate students who are *reluctant* to speak English?
- How do you maintain *authority* in the classroom?

After the interview task, learners carried out a second sentence-completion activity with 10 items that were based on the questions that they asked during the role-play task. Once again, the target words were embedded in sentence-medial position. Learners were asked to complete the sentences based on their interview notes (see Example 3).

Example (3)

- He motivates *reluctant* students to speak English by________________________

  ____.

- He maintains *authority* in the classroom by________________________________

  ____.

Finally, for the last step in the task sequence, learners carried out an information exchange and consensus task. Learners were asked to share with each other their interview notes from the second sentence-completion activity and discuss whether they would like to hire the applicant for the position. There were two versions of the handouts given for the two sentence-completion activities and the interview task questions. In other words, each learner in a dyad had a different set of 10 sentences/questions. The purpose of providing learner dyads with different target words and sentences was to eliminate the possibility of a priming effect. Furthermore, working on different sentences gave learners the opportunity to share different ideas with each other, which made the information exchange and consensus tasks more meaningful.

5.1.4. Target words

There was a total of 20 target words that were selected from the academic word list (Coxhead, 2000), 10 three-syllable and 10 four-syllable words with primary stress on the second syllable. These words were divided into two sets of 10 words with 5 three-syllable and 5 four-syllable words. As learners carried out the communicative tasks in pairs, the first set of target words were embedded into the handouts that were given to Learner A and the second set of target words were embedded into the handouts that were given to Learner B. This way, the target words were counter balanced. Also, learners did not have the opportunity to hear the pronunciation of the target words from their partner prior to producing those words themselves. The words that were used for the pilot study can be seen in Table 5.1.

Table 5.1

Target words used in the pilot study

Set 1	component	objective	exposure	assessment	encounter
	approachable	communicate	authority	participate	appreciate
Set 2	establish	potential	explicit	dynamic	reluctant
	philosophy	evaluate	alternative	analogy	significant

5.1.5. Procedure

The learners were randomly assigned to FTF-recast (*n* = 13), SCMC-recast (*n* = 16), FTF-control (*n* = 15), and SCMC-control groups (*n* = 13). Originally, 16 learners were randomly assigned to each group; however, after the removal of data, the distribution patterns slightly differed. Learners attended data collection sessions in pairs in a quiet room located within the Intensive English Program premises. At the onset of data collection, learners were informed about the procedure and they were given a chance to ask questions about the procedure or any word that they did not understand. When learners asked about the meaning of a word, which only happened occasionally, they were provided with a definion which did not include the pronunciation of the target word in order to prevent a priming effect. After filling out the background survey, learners carried out the first sentence-completion activity for which they were instructed to complete the open-ended sentences related to teaching and learning English based on their opinion and experience. They were ensured that there were no right or wrong answers. After the first sentence-completion activity, learners carried out the information exchange task by taking turns to read their sentences to their partner and to comment on the sentences shared by their partner. Next, learners carried out the interview task with faculty members from the same university who were a native speaker of Canadian English, a native speaker of British English, and a near-native speaker of English (the researcher). Prior to data collection, both native speakers had been informed about the procedures of the study and they had been trained on providing recasts. The FTF and SCMC intervention groups were randomly assigned to interview one of the two native speakers. On the other hand, the FTF and SCMC control groups interviewed the researcher. Both the FTF and the SCMC groups attended the data collection session in pairs in the same quiet room; however, the SCMC intervention and control groups' interactions with the interviewee took place over Skype using a MacBook Pro computer. The computer speakers were tested prior to each data collection session to ensure that learners did not experience any difficulty with hearing the interviewee. Learners in the intervention group received corrective feedback in the form of a recast when they mispronounced a target word by placing the primary stress on the wrong syllable. Learners in the control group did not receive any form of corrective feedback. Once the interview task was completed, learners were given the second sentence-completion activity and asked to complete the sentences based on the responses they received from the interviewee. For the last step of the task sequence, which was the information exchange and consensus task, learners shared their interview notes with each other by reading aloud their sentences and commenting on the notes shared by their partner. Then, they decided whether they should hire the candidate or not. Finally, learners filled out the exit survey. The entire data collection session lasted approximately one hour. The pretest data was collected during the first information exchange task and the posttest data was collected during the second information exchange and consensus task. During the communicative tasks, learners wore Shure WH20 XLR brand headworn microphones that were connected to Roland Duo-Capture EX brand USB audio interface. The unidirectional microphone was positioned at 30 degrees off-axis and about 3 cm away from learners' mouth. The speech files were saved onto an iPad Pro as 44,100 Hz .wav files.

5.2. Acoustic analysis

5.2.1. Coding

Acoustic analyses were carried out on Praat (Boersma & Weenink, 2018). Prior to coding, unintelligible tokens were removed. The syllable boundaries for the remaining tokens were determined using visual and auditory information. Target words were extracted from longer stretches of speech data using a Praat script (Kirkham, 2014). After that, another Praat script (Kirkham, 2015) was used to extract duration, pitch, and intensity measures from each syllable. The duration values that were originally in miliseconds were multiplied by 1000 and converted to seconds. The intensity values were maintainted in decibels. Finally, the pitch values were maintained in hertz. Before carrying out statistical analyses, subtraction ratios were calculated for each acoustic correlate. In the case of three-syllable words, the subtraction ratios were calculated by subtracting the sum of the first and third syllables from the second syllable. For the four-syllable words, the sum of the first, third, and fourth syllables were subtracted from the second syllable. Below are the formulas used for calculating subtraction ratios.

$$\text{three-syllable words: } S_{norm} = S2-(S1+S3)$$

$$\text{four-syllable words: } S_{norm} = S2-(S1+S3+S4)$$

5.2.3. Mixed-effects models

Following the research design, a mixed-effects model was fitted using the lme4 package version 1.1-8 (Bates, Mächler, Bolker, & Walker, 2015) with condition, modality, and time as fixed effects. Learners and words were entered into the model as random effects. A second mixed-effects model was created with the addition of syllable as a fixed effect. The addition of syllable as a fixed effect improved the model in the case of duration measures, $\chi2(8) = 18.6$, $p = .02$, but it did not have any effect in the case of intensity measures, $\chi2(8) = 2.91$, $p = .94$, or pitch measures, $\chi2(8) = 9.75$, $p = .28$. Based on model comparisons, syllable was maintained as a fixed effect for the analysis of duration but it was removed from the analysis of pitch and intensity. Pairwise comparisons were conducted using the multcomp package version 1.4.1 (Hothorn, Bretz, & Westfall, 2008).

5.2. Results of acoustic analyses

To being with duration, the results showed that there were no significant differences between the pretest and posttest productions of duration by the FTF recast group regardless of whether the target words had three syllables (ß = 153.61, SE = 79.04, z = 1.94, p = .24) or four syllables (ß = −163.80, SE = 110.84, z = −1.48, p = .52). Similarly, there was no statistical improvement in duration productions by the

FTF control group either for three-syllable words (ß = −106.06, *SE* = 56.25, $z = -1.89$, $p = .27$) or four-syllable words (ß = 114.96, *SE* = 79.58, $z = 1.45$, $p = .55$). When it comes to the SCMC groups, the SCMC recast group did not exhibit any gains in duration for three-syllable words (ß = 11.66, *SE* = 55.12, $z = 0.21$, $p = 1.00$) or four-syllable words (ß = −62.60, *SE* = 77.38, $z = 0.81$, $p = .94$). Finally, there were no significant differences in duration between the pretest and posttest productions of three-syllable words (ß = 10.12, *SE* = 24.93, $z = 0.41$, $p = .99$) or four-syllable words by the SCMC control group (ß = 58.28, *SE* = 84.73, $z = 0.69$, $p = .97$).

The results did not indicate any improvement for the intensity measures either. FTF recast group had slightly higher gains on the posttest (ß = 2.01, *SE* = 2.77, $z = -0.72$, $p = .84$) compared to the FTF control group (ß = −1.62, *SE* = 1.99, $z = -0.82$, $p = .79$). On the other hand, the SCMC recast group had slightly lower intensity values on the posttest (ß = −1.79, *SE* = 1.94, $z = -0.92$, $p = .71$) compared to the SCMC control group (ß = 0.51, *SE* = 0.83, $z = -0.63$, $p = .90$).

As it was the case for duration and intensity, learners did not show any statistical improvement in their production of syllable pitch. The pitch productions by the FTF recast group were higher on the posttest (ß = 5.56, *SE* = 12.93, $z = 0.43$, $p = .97$), whereas for the FTF control group, the pitch productions were lower on the posttest (ß = −10.86, *SE* = 9.25, $z = -1.17$, $p = .54$). When it comes to the SCMC groups, the SCMC control group produced higher pitch values on the posttest (ß = 6.59, *SE* = 3.83, $z = 1.72$, $p = .23$), whereas the SCMC recast group produced lower pitch values (ß = −8.27, *SE* = 9.03, $z = -0.92$, $p = .72$).

5.3. Auditory analyses

In addition to acoustic analyses, the researcher conducted auditory analysis of the data together with a native-speaker of British English who had a background in

Teaching English to Speakers of Other Languages. The British NS rated a subset of 100 tokens. He was informed about the procedure and completed a brief training session prior to rating the data set. Kappa analysis yielded a strong agreement between the two raters ($\kappa = .86$, $p = .00$).

According to rater judgments, the SCMC recast group correctly placed lexical stress 40.16% of the time on the pretest, and they were able to increase the percentage of correct placement to 47.75% on the posttest. When it comes to the FTF recast group, they correctly placed lexical stress 37.63% of the time on the pretest and increased their correct production to 59.26% on the posttest. In short, the FTF group showed higher gains based on audiotory analysis. Based on the auditory analysis, the researcher was able to identify 20 tokens that were marked as inaccurate on the pretest but as accurate on the posttest. These were all three-syllable words produced by participants from FTF and SCMC recast groups. In order to analyze how the acoustic correlates of these 20 words changed over time, a new mixed-effects model was fitted using time as a fixed effect, and learners and words as random effects. The results showed a significant increase in syllable duration on the posttest ($\beta = 92.27$, $SE = 30.11$, $t = 3.07$, $p = 0.00$). However, the changes in pitch ($\beta = 10.27$, $SE = 7.58$, $t = 1.36$, $p = 0.19$) and intensity measures ($\beta = 1.72$, $SE = 1.38$, $t = 1.25$, $p = 0.23$) were not statistically significant.

5.4. Discussion and changes for the main study

Although the acoustic analyses did not yield statistical gains for any of the correlates, certain patterns that emerged during the pilot study informed the changes made for the main study. For example, the inclusion of syllable as a fixed effect improved the model when the dependent variable was duration. What is more, despite lack of statistical significance, the results showed that the highests gains in duration were achieved in the case of three-syllable words. Further analysis of 20 words that were auditorily identified as inaccurate on the pretest but accurate on the posttest showed that duration was the only acoustic correlate that improved significantly on the posttest. Although it is difficult to provide a specific reason for why three-syllable words exhibited more gains, one possible reason could be the combination of the developmental readiness of the learners and the requirements of the communicative tasks that they carried out. Therefore, it was decided that the main study should mainly focus on three-syllable words.

Another issue was statistical power. In the pilot study, modality was a variable, which necessitated having four experimental groups. However, this led to low statistical power because only around 15 learners were placed in each group due to logistical reasons. The number of learners that the researcher could recruit was not going to be more than 70 learners per academic year. In order to tackle this issue and increase statistical power for the main study, modality was taken out of the study design. This way, it was possible to assign learners into two groups (intervention and control) and have around 35 learners in each group, which would lead to higher statistical power.

The calculation of relative difference measures has also been improved for the main study. In the pilot study, a subtraction ratio was used to calculate a single number that represents the value of the second syllable relative to the other syllables in the same word. After that, this number was used when analyzing the differences between pretest and posttest productions of the three acoustic correlates. However, upon coming across an article by Taylor and Wales (1987), which makes a case for using the contrast ratio, I decided to use the contrast ratio when analyzing the data for the main study. The contrast ratio is obtained by subtracting the sum of unstressed syllables from the stressed syllable and dividing that number by the sum of all syllables. In other words, compared to the subtraction ratio, it includes one more calculation, which is diving the number obtained from the subtraction by the sum of all syllables. Taylor and Wales (1987) explain that a contrast ratio is a robust predictor of accent compared to the subtraction ratio and provide evidence in the form of a series of analyses of the same data using different ratios. A more detailed of discussion on this choice is provided in Chapter 6.7.1, where coding of the data for the main study is discussed in detail. In short, the second improvement was using the contrast ratio instead of the subtraction ratio when obtaining the number that represents the value of the second syllable relative to other syllables in the same word.

One final improvement for the main study was the method employed for extracting pitch measurements. For the pilot study, the pitch measurements were directly extracted from each syllable using the default values of 75 Hz for the pitch floor and 600 Hz for the pitch ceiling. However, in order to improve the robustness of the analyses, a parameter optimization method recommended by Keelan, Lai, and Zechner (2010) was used for the main study. This optimization method is requires calculation of individual pitch floor and pitch ceiling values prior to extracting pitch measurements. The authors show that the optimization method substantially improves the accuracy of pitch extraction algorithms such as Auto Correction (Boersma, 1993), Cross Correlation (Atal, 1968), and Sawtooth Waveform Inspired Pitch Estimator (Camacho, 2007). Since Praat uses Auto Correction, it was decided that the optimization method recommended by the authors be implemented for the main study. The details of how the optimization was carried out is explained in Chapter 6.7.1.

CHAPTER 6: METHODS AND STUDY DESIGN

6.1. Research questions

The current study explores the impact of recasts as a form of corrective feedback on the development of primary stress in L2 English. The findings from the pilot study, which were reported in Parlak and Ziegler (2017), showed that recasts on non-target-like productions of primary stress have the potential to trigger a change in immediate posttest production in the form of increased syllable duration. The findings confirmed previous studies that emphasized the reliance on syllable duration by L1 Arabic speakers when they produce lexical stress in English. As a follow-up study, the current study mainly focuses on three-syllable words as the target vocabulary and investigates the impact of recasts provided during face-to-face interaction. The goals of the current study were to investigate the relationship between recasts and the developmental changes in the acoustic correlates of primary stress in English produced by L1 Arabic speakers and whether those changes would affect listener perceptions. In light of these goals, the research questions below were formulated:

RQ1: To what extent is providing recasts on learners' primary stress errors during a communicative activity associated with more target-like production of primary stress?

If so,

RQ2: *How* do recasts impact learners' acoustic realization of primary stress?

RQ3: *How* do recasts impact listener judgments of learners' stress placement?

6.2. The context

The study was conducted in the United Arab Emirates (UAE). The UAE attracts immigrants from all parts of the world due to its fast-growing economy that is mainly fueled by petroleum, natural gas, tourism, and trade. The perpetual immigration and multinational fabric of the country necessitates the use of English a lingua franca for everyday communication that occurs in public spaces as well as in the fields of education and business. As such, attaining good English language skills is critical for speakers of other languages who wish to pursue higher education and build a career in the UAE.

There are a number of English-medium universities in the UAE that provide western-style education and some of them are based on the American model. The current study was conducted at the Intensive English Program of one of those American universities. The mission of the program is to prepare students who have been admitted to the university but have not fulfilled the English language proficiency requirement for matriculation into their majors. The program accepts students who have scored between 32 and 79 on the internet-based Test of English as a Foreign Language (TOEFL iBT), between 400 and 547 on the paper-based TOEFL, or between 4.5 and 6.0 on the International English Language Testing System (IELTS). Completion of the program requires fulfilling the coursework requirements successfully, passing the in-house exit exam, and receiving a score of 80 (550) or above on the TOEFL, or 6.5 or above on the IELTS.

6.3. Participants

There were two main groups of participants; learners and listeners. The first group was L1 Arabic speakers studying English at the Intensive English Program. The second group was faculty members who spoke English as a native language and taught undergraduate-level academic writing courses or English literature courses at the same university. The first group participated in the language activities and henceforth will be referred to as *learners*. The second group carried out a listener perception experiment and henceforth will be referred to as *listeners*. Prior to participant recruitment, ethical approval was received from Lancaster University and subsequently a second ethical approval was received from the institutional review board of the university where the study was conducted.

6.3.1. Learners

Learners were recruited through classroom announcements. Prior to class visits, the Program Director of the Intensive English Program and the teachers were informed about the goals of the study. The teachers were also ensured that the study would not have a negative impact on learners' attendance or classroom performance, as data collection sessions were to be scheduled during learners' free time. During classroom announcements, learners were given brief information about participation in the study which was followed by a quick question-and-answer session. Then, a sign-up sheet was passed around for learners who were interested in the study. The researcher contacted the learners who signed up for the study via e-mail and Whatsapp (a mobile messaging application) to arrange a time for data collection. Each learner was given a 15-dirham gift voucher (approximately 3 British Pounds at the time of the study) for a cup of coffee and a donut that they could redeem at the university cafeteria. Considering the socio-economic background of the learners, the 15-dirham gift vouchers were nothing more than a simple token of appreciation of their participation.

Learners were 1 Hindi and 73 Arabic speakers of L2 English (44 males, 30 females). According to the background survey they completed at the outset of data collection (see Appendix B), the dialects of Arabic spoken by learners were Gulf (N = 39), Levantine (N = 22), Egyptian (N = 10), and Sudanese (N = 2). Figure 6.1 provides a graphical view of the frequency distribution of the Arabic dialects spoken by learners. One way to group Arabic dialects is dividing them into western dialects and eastern dialects. The former includes Moroccan, Algerian, and Tunisian dialects, whereas the latter includes Egyptian, Levantine, Gulf, and Sudanese dialects (Mustafawi, 2018). These two groups of dialects have different prosodic patterns particularly. One of the main reasons for prosodic differences is that western dialects have reduced short vowels, which leads to differences in rhythm (Barkat, Ohala, & Pellegrino, 1999; Ghazali, Hamdi, & Barkat, 2002). All learners in the current study were speakers of eastern dialects who speak with similar prosodic patterns. In fact, stress placement patterns in eastern dialects also follow the same pattern apart from minor exceptions (Mustafawi, 2018). Therefore, their data was grouped together. The data from the Hindi speaker was removed as Arabic was not her L1. In addition, the data from five other learners had to be removed either due to technical problems with the recordings or problems with pronunciation that affected their intelligibility severely. After the removal of their data, there were 68 remaining learners, 32 of whom were randomly placed into the control group (21 males, 11 females) and 36 in the intervention group (21 males, 15 females).

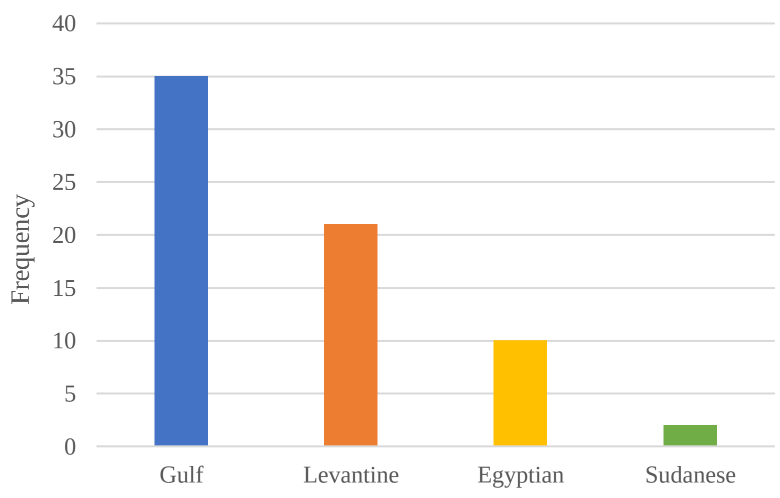


Figure 6.1 Dialects of Arabic spoken by learners

Learners' proficiency levels were determined based on the TOEFL (*N* = 19) and IELTS (*N* = 49) scores they reported as part of their responses to the background survey. The TOEFL scores were converted to IELTS scores for purposes of consistency and comparability by following the conversion tables published by Educational Testing Service (2010). Learners' proficiency levels ranked mainly between a score of 5.5 and 6 (*N* = 58) on the IELTS. There were five learners with a score of 5, four learners with a score of 6.5, and one learner with a score of 7. However, it is important to note that the learners who reported a score of 5 also mentioned that they had taken the IELTS two to five months prior to the date of data collection. As they had been studying English for a couple of months in the same classrooms with other learners who scored 5.5 on the IELTS more recently, it is assumed that their proficiency level was not dramatically different from that of their classmates. On the other hand, the five learners who reported a score of 6.5 or 7 mentioned that they took the IELTS one to four weeks prior to data collection. As the pilot study showed, learners' proficiency levels were appropriate for carrying out the language activities.

Learners' mean age was 18 (SD = 0.86). Eight of them stated that they spoke a L2 in addition to English (a local Sudanese dialect, Hindi, Turkish, Farsi, French, Russian, and Italian). Three learners reported that they had received some form of pronunciation instruction in the past, which was delivered as part of a speaking and listening class. As the UAE is a context where English is used as a lingua franca, learners were also asked to indicate their percentage of English use by three types of interlocutors: non-native speakers who are not family members, non-native speakers who are family members, and native speakers. Based on the self-reported numbers, learners used English to communicate with non-native speakers (Mpercentage = 7.07, SD = 13.33 with non-native speakers who are family members; Mpercentage = 53.19, SD = 23.50 with non-native speakers who are not family members) more than they did with native speakers (Mpercentage = 39.74, SD = 21.04). Five learners had lived in an English-speaking country prior to data collection with their length of residence ranging from two months to two years. Finally, one learner reported that he has difficulty pronouncing /ɹ/. The rest of the learners reported no hearing or speech related problems. It should be noted that the five learners who have lived abroad or the learner who reported difficulty with pronouncing /ɹ/ did not display any unusual patterns of production throughout the data collection. Table 6.1 provides a summary of the demographic details of the 68 learners.

Table 6.1

Learner demographics by group

		Control Group	Intervention Group
Age (years)		17.78 (*SD* = 0.75)	18.19 (*SD* = 0.92)
Gender	Male	21	21
	Female	11	15
Arabic Dialect	Gulf	15	20
	Levantine	10	11
	Egyptian	5	5
	Sudanese	2	0
IELTS Scores	5.5 – 6.0	31	32
	6.5 – 7.0	1	4

6.3.2. Listeners

After preparing the speech data collected from learners for the listener judgment experiment, native-speaker listeners were contacted during their office hours. The goals of the study and their role, if they chose to participate, were explained. As payment for their participation, listeners were given a 50-dirham gift voucher (approximately 10 British Pounds at the time of data collection) that they could redeem at various retail stores across the UAE.

Listeners were 10 native speakers of North American English (5 males, 5 females) teaching undergraduate-level courses. 9 of them taught academic writing courses and 1 taught English literature courses. Their mean age was 48.6 (*SD* = 8.46). According to the background survey (see Appendix C), one listener spoke Urdu, one Italian, one Greek, and one Spanish as a heritage language. The listener who spoke Spanish as a heritage language also spoke French as a foreign language. At the time of data collection, listeners' length of residence in an Arab country, including any other Arab country they had lived in prior to coming to the UAE, ranged between 4 to 20 years. However, it is important to highlight that they lived in country and worked at an institution that allowed them to be constantly in touch with other native-speakers of English, including speakers of North American English, on a daily basis. What is more, they visited their home country at least once a year for a couple of weeks. Nine of them indicated that they were very familiar with L2 English spoken by L1 Arabic speakers. Considering the amount of time that they had spent in Arabic-speaking countries at the time of data collection and their level of familiarity with the type of English spoken by L1 Arabic speakers, listeners who took part in the current study can be considered experienced listeners, where experience refers to the amount of exposure to the specific variety of L2 speech (Kennedy & Trofimovich, 2008). One listener mentioned that his hearing was fine but at the same time not what it used to be because of "too much loud Rock 'n' Roll as a youth." A summary of listener demographics is provided in Table 6.2.

Table 6.2

Listener demographics

Age (years)		48.6 (*SD* = 8.46)
Length of Residence in an Arab Country (years)		10.6 (*SD* = 5.27)
Gender	Male	5
	Female	5
English Dialect	North American	10
Heritage and Second Languages	Urdu	1
	Greek	1
	Italian, French	1

6.4. Instruments

There were four main sets of instruments: the language tasks that learners carried out, the target words that were embedded into the language tasks, the exit survey that learners filled out at the end of data collection, and the listener judgment experiment set up for triangulation purposes.

6.4.1. Language tasks

A detailed explanation of tasks and pre-task activities has been provided earlier in Chapter 5.1.3. The tasks and the pre-task activities used for the current study were mainly the same. However, as the current study focused on three-syllable words, the sentences that were embedded with four-syllable words in the pilot study had been modified by embedding three-syllable words. This way, each learner in a dyad had a set of 10 three-syllable words to produce. The procedure for carrying out the task sequence was also kept the same. The complete set of handouts given to learners for the current study can be seen in Appendix A.

6.4.2. Target words

The three handouts that were given to learners during the first sentence-completion activity, the interview task, and the second sentence-completion activity contained the carrier sentences and questions that were embedded with the target words. Carrier sentences are used commonly to elicit speech data in phonology and phonetics studies (e.g., Colantoni, Marasco, Steele, & Sunara, 2014; McAllister, Flege, & Piske, 2002) as well as recent interactionist studies that focused on development of L2 phonology (e.g., Saito & Lyster, 2012a, 2012b). They can be particularly useful as they allow researchers to ensure that the target features they are interested in studying are elicited consistently. Also, through the use of carrier sentences, researchers are able to control the phonological environment, such as preceding and following vowels/consonants, as well as the non-phonological environment, such as morphological, lexical, and syntactic elements surrounding the target language feature (Colantoni et al, 2015). In their review of methodological issues in studies focusing on word stress, Roettger and Gordon (2017) highlight the importance of controlling for context by avoiding target words produced in isolation and placing them at a point in the carrier sentence away from phrasal boundaries, as words placed at phrasal boundaries can lead to a confounding effect with boundary-associated tonal effects. They also argue that isolated words can make it difficult to separate word stress from phrase-level prominence, which was referred to as pitch-accent earlier in Chapter 4.2. Considering these recommendations, the target words were embedded in sentence-medial position for each carrier sentence. Another motivation to use carrier sentences was to ensure consistency of eliciting the same target features from each participant. How a teacher or a researcher envisions task-as-workplan, defined as expectations from the task, can be different from task-in-process, defined as what actually happens during the execution of the task, as learners may have their own way of interpreting and carrying out the task (Breen, 1987, 1989; Samuda & Bygate, 2008; Seedhouse, 2005). Considering this possibility, the carrier sentences were also instrumental in minimizing the differences between task-as-workplan and task-in-process, which ensured consistently in data collection.

A potential disadvantage of using read materials compared to spoken language could be that participants behavior during controlled speech may not be the same as their behavior during naturally produced speech. As a result, one may argue that research on pronunciation would yield better results, or findings would be more meaningful when speech samples are naturally produced. It is hard to deny the ecological validity of spontaneous and naturally produced language. However, spontaneous speech also has an important limitation from the research perspective, which is the lack of control over target language features. In many cases, lack of control makes it difficult to collect enough data on a specific language feature and run statistical analyses. Furthermore, in an article that addresses criticism directed toward lab speech, Xu (2010) debunks several misconceptions about lab speech that consider it slow, unnatural, monotonous, and lacking communicative functions. Xu explains why these misconceptions are unfounded and explains that lab speech allows researchers to use theory-based approaches and increases generalizability of the results as it minimizes the number of uncontrolled factors. In short, read materials in the form of carrier sentences are still a useful method for collecting speech data, especially when subjects have the level of literacy that would allow them to read in the target language, which was the case for the current study. It is also important to highlight that the carrier sentences in the current study were used as part of communicative tasks that resulted in highly interactive speech data, which is different from lab studies in which participants are given a particular carrier phrase and are asked to read it multiple times by substituting the target word for each iteration (e.g., Say X quickly).

The choice of target words embedded into the carrier sentences was partially informed by the findings from the pilot study. The learners who participated in the pilot study (Parlak & Ziegler, 2017) exhibited more acoustic gains for their placement of primary stress in the case of three-syllable words compared to four-syllable words. This was perhaps an issue related to developmental readiness (Han, 2002; Mackey & Philp, 1998) in combination with the requirements of the tasks. Because the learners in the current study were sampled from the same population, it was decided that focusing only on three-syllable words would be more level-appropriate. As a result, the current study focused on 20 three-syllable words with primary stress on the second syllable as the target vocabulary. There were two additional considerations that influenced the selection of target words, namely orthographic transparency and syllable structure. Orthographic transparency was taken into consideration as learners were required to read aloud carrier sentences with the target words embedded in them. In general, the English writing system has low orthographic transparency because there is no one-to-one mapping between the phonemes and the orthographic forms. This could potentially create a challenge for L2 speakers as previous research has shown that orthographic input may create non-target-like phonological representations in learners' mental lexicon (Basetti, 2006). It may also lead to non-target-like production due to issues with voicing (Young-Scholten & Langer, 2015), epenthesis

(Basetti & Atkinson, 2015), or pronunciation of silent letters (Aro & Wimmer, 2003; Browning, 2004). In order to mitigate the possible negative effects of orthography on learners' pronunciation, the target words selected for the current study did not have any irregularities or silent letters. Moreover, the syllable structure of each word was closely tied to its orthographic form. When L2 learners have difficulty pronouncing consonant clusters, they may employ epenthesis in order to maintain the consonants presented in orthographic form (Young-Scholten, 1998). Therefore, complex syllable structures, in particular long consonant clusters, were avoided. Altogether, there were seven patterns of syllable structures: CV-CV-CVCC ($N = 1$), CV-CVC-VC ($N = 6$), CVC-CV-CVC ($N = 1$), CVC-CVC-VC ($N = 3$), CVC-CCV-CVC ($N = 1$), CVC-CVC-CVC ($N = 6$), and CVC-CVC-CVCC ($N = 2$).

The twenty words were divided into two sets with ten words in each set. The words from Set 1 were embedded into the handouts that were given to Student A and words from Set 2 were embedded into the handouts that were given to Student B[1]. The two sets were created in order to counterbalance the target words and to avoid production of the same target word by both participants, which could have led to a priming effect. The complete list of target words can be seen in Table 6.3.

Table 6.3

Target words

Set 1	dynamic	develop	revision	consider	suspicion
	confusion	perception	responsive	contention	consistent
Set 2	position	duration	diminish	formation	condition
	compassion	companion	consensus	convention	persistent

[1] The word *student* was used on the handouts instead of the word *learner*.

6.4.3. Exit survey

Upon completion of the language tasks, learners were given an exit survey (see Appendix D). Mackey and Gass (2016) recommend that researchers consider four goals when designing a survey: “simple, uncluttered formats; unambiguous, answerable questions; review by several researchers; piloting among a representative sample of the research population” (p. 105). Additionally, it is important to keep the survey at a reasonable length as unnecessarily long surveys may adversely affect the results (Dörnyei, 2007). Following these guidelines, the exit survey was created using Qualtrics (2018). The exit survey consisted of likert-scale and open-ended items formulated to elicit responses that would give clues about whether learners in the intervention group directed their attention to pronunciation and whether they noticed the recasts that were provided during the interview task.

6.4.4. Listener judgment experiment

Derwing and Munro (2015) suggest that pronunciation gains observed through acoustic measures may not always correspond to a change in listener perceptions. Therefore, they highlight the importance of analyzing listener judgments in addition to acoustic measures when studying L2 speakers’ pronunciation gains (Derwing & Munro, 2009, 2013). Following their recommendation, a listener judgment experiment was set up on PsychoPy (Peirce, 2009) for triangulation purposes.

The experiment contained randomizations of 120 words, which were pretest and posttest productions of 30 words from the control group and 30 words that received a recast from the intervention group. The randomizations of 120 words were equally divided into two sets to allow listeners to take a short break halfway through the rating process, in case they felt the need for it. A three-point nominal scale was used to collect the listener perception data with the labels *Syllable 1*, *Syllable 2*, and

Syllable 3. Listeners were required to indicate which syllable they thought carried the primary stress upon hearing a word. The software allowed listeners to listen to target words multiple times and recorded their response automatically before presenting the next word. Further details on the listener experiment are provided in Chapter 6.6.2.

6.5. Study design

The study design can be seen in Figure 6.2. The study followed a pretest-intervention-posttest design. Learners attended the data collection sessions in pairs. They carried out the sentence-completion activities individually; the information-exchange and decision tasks in pairs; and for the interview task, they interviewed the researchers in pairs. The learners in the intervention group received a recast only as a response to their production of the target words with a non-target-like stress pattern. They did not receive any form of feedback when they mispronounced other words. No feedback was provided for the learners in the control group.

Four key points highlighted by previous research were taken as a guide when operationalizing the provision of recasts. The first point is that perceptual salience of recast may be enhanced by providing the target form in isolation immediately after the non-target-like production which allows learners, whether they are learning their first language as a child or second language as a child or an adult, to juxtapose their production with the target form (Carpenter et al., 2006; Doughty, 2001; Farrar, 1990; Long, 1996, 2007; Long, Inagaki, & Ortega, 1998; Saxton, 1997). Secondly, utilizing prosodic emphasis when providing a recast could help with enhancing its salience (Loewen & Philp, 2006), although the relationship between recasts enhanced with prosodic emphasis and target-like responses is still under discussion (McDonough, Crowther, Kielstra, & Trofimovich, 2015). Thirdly, the length of a recast may have a direct impact on learners' processing capacity. As such, shorter recasts are deemed

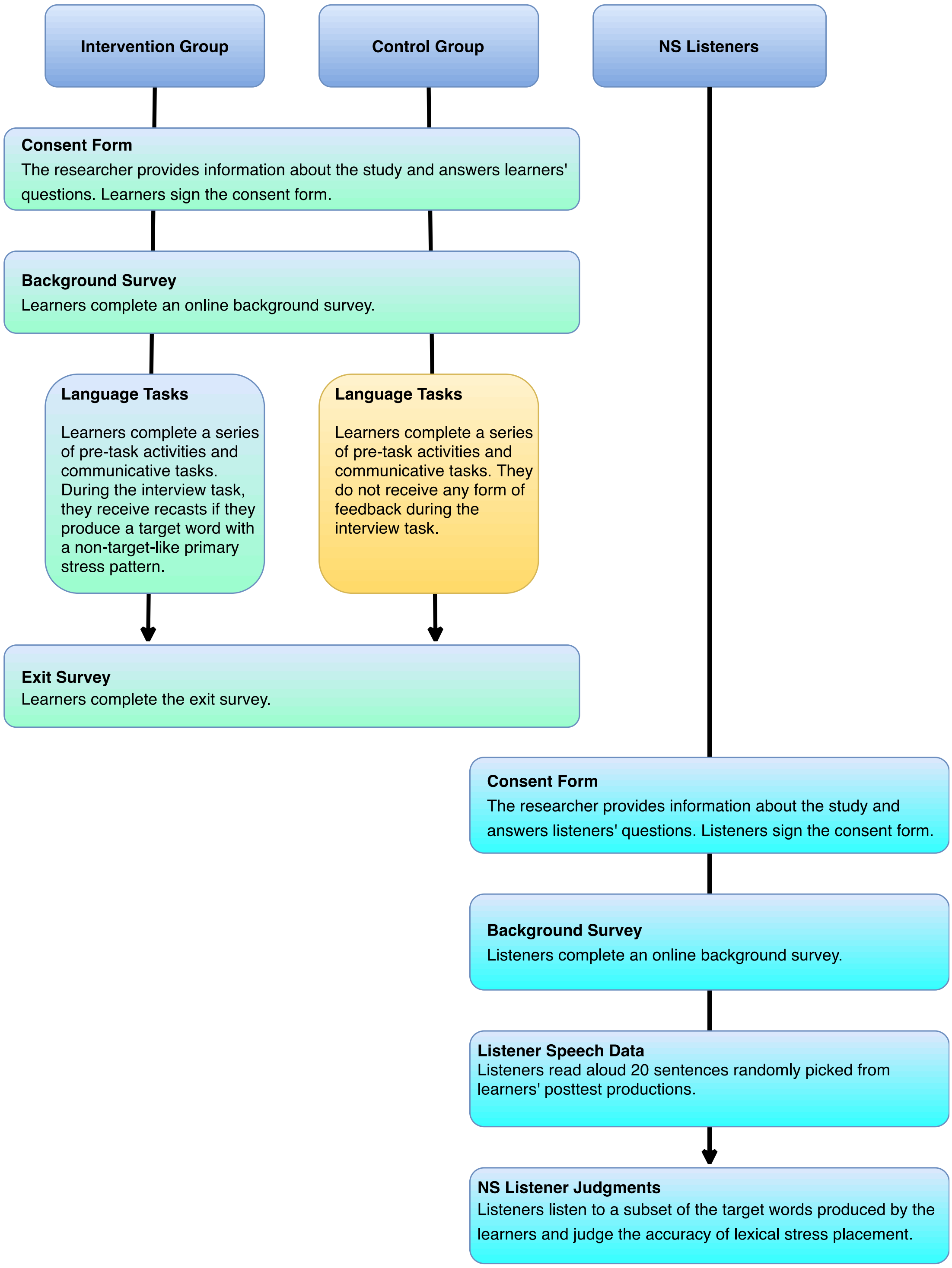


Figure 6.2 Study design

potentially more useful because they are less demanding in terms of working memory load (Philp, 2003; Sheen, 2006). Finally, consistent focus of recasts and their intensity can increase the salience of recasts, and therefore, be more useful for learners (Han, 2002; Mackey & Philp, 1998). The current study followed a two-step recast provision method guided by these four key points.

First, the target word was repeated in isolation by placing extra emphasis on the correct syllable as a response to learners' non-target-like production. This first recast was very short in length as it was only a single word, carried additional prosodic emphasis on the second syllable, and provided an opportunity for learners to juxtapose the two different productions of the same word. Immediately after that, the target word was repeated as part of the response given to the question asked by the learner, again with increased prosodic emphasis on the second syllable.

The second recast was intended to reinforce the effect of the first recast while maintaining the natural flow of the conversation. According to the terminology suggested by Sheen (2006), the recasts used in the current study were declarative, single-word recasts with one change that were repeated once in isolation and once incorporated into a response immediately following the isolated recast.

Another consideration was controlling for modified output as it has been done in previous studies (e.g., Goo, 2012; Leeman, 2003) and recommended by Yilmaz (2016). Modified output was not a target of investigation; however, in order to control for possible effect of modified output and to maintain consistency across different instances of feedback provision, learners were not given an opportunity to produce modified output. After the initial isolated recast, the researcher continued with his response and answered the question asked by the learner. To prevent opportunities for modified output, the researcher utilized fillers and pauses to hold the turn and link the isolated recast to the complete response in a natural way (see Example 4, 5, and 6).

Example (4)

Learner: Aaa… how do you make sure that you are consisTENT when grading exams?

Researcher: Hmmm… well… conSIStent… how do I make… that's a very good question. To be conSIStent, which is very important I think, I do not look at their names when I grade their papers.

Example (5)

Learner: How do you motivate students to be persisTENT when they face difficulties?

Researcher: perSIStent… well… I think, again, it is important to help students become perSIStent. What I do is… I tell them that I am always there to help them, if they have any problems.

Example (6)

Learner: Ok… umm… do you consiDER students' interests when creating lessons?

Researcher: umm… conSIDer… yes… aaa… I conSIDer students' interests because I want to make the lessons interesting for them.

In some cases, learners interjected with a brief confirmation such as "yeah" or "hmm", which indicated that they were actively and naturally participating in the interaction (see Examples 7, 8, 9 and 10).

Example (7)

Learner: What is the best COMpanion of a language learner?

Researcher: ummm, comPANion..

Learner: Yeah

Researcher: The best comPANion of a language learner is, in my opinion, a dictionary.

Example (8)

Learner: Does having com… COMpassion make a teacher better at their job?

Researcher: comPASSion…

Learner: mmm

Researcher: Yes, if a teacher has comPASSion, they will be good at their job.

Example (9)

Learner: How do you deveLOP good relationships with your students?

Researcher: Ummm… deVELop…

Learner: Hmm…

Researcher: To deVELop good relationships, well, I make myself available to them.

Example (10)

Learner: What is the ideal… duraTION for a language class?

Researcher: duRAtion…

Learner: Yes

Researcher: The ideal duRAtion would be… I think one hour is good.

6.6. Procedure

The data was collected in two main parts. First, the learner data was collected and coded. After that, a portion of the learner data was used to collect the perception data from native-speaker listeners.

6.6.1. Data collection from learners

Prior to data collection, learners were randomly assigned to the control group or the recast group. They attended the data collection session in pairs as the language tasks were designed for pair work. The data was collected in a quiet study room located within the Intensive English Program premises. First, learners were informed about the procedure and then they were given time to read the consent form and ask questions about the study. After signing of the consent form, learners were asked to complete the background survey that was created by using Qualtrics (2018). Learners completed the survey on a laptop computer provided by the researcher. When a learner did not understand a survey item, a brief explanation was provided. Before learners began carrying out the pre-task activities and communicative tasks, they were asked to set their mobile phones to flight mode in order to prevent possible electromagnetic interference on the recording equipment as well as to ensure disruption-free data collection sessions.

Next, the procedure shown in Figure 6.3 was followed. Learners were given the handout for the first sentence-completion activity, which required them to complete open-ended sentences about language learning and language teachers. They were informed that there were no right or wrong answers and that they could draw from their personal experience when completing the sentences. They were also asked to limit their response for each item on the handout to a single sentence. When learners did not understand a sentence or a target word, the researcher provided a brief

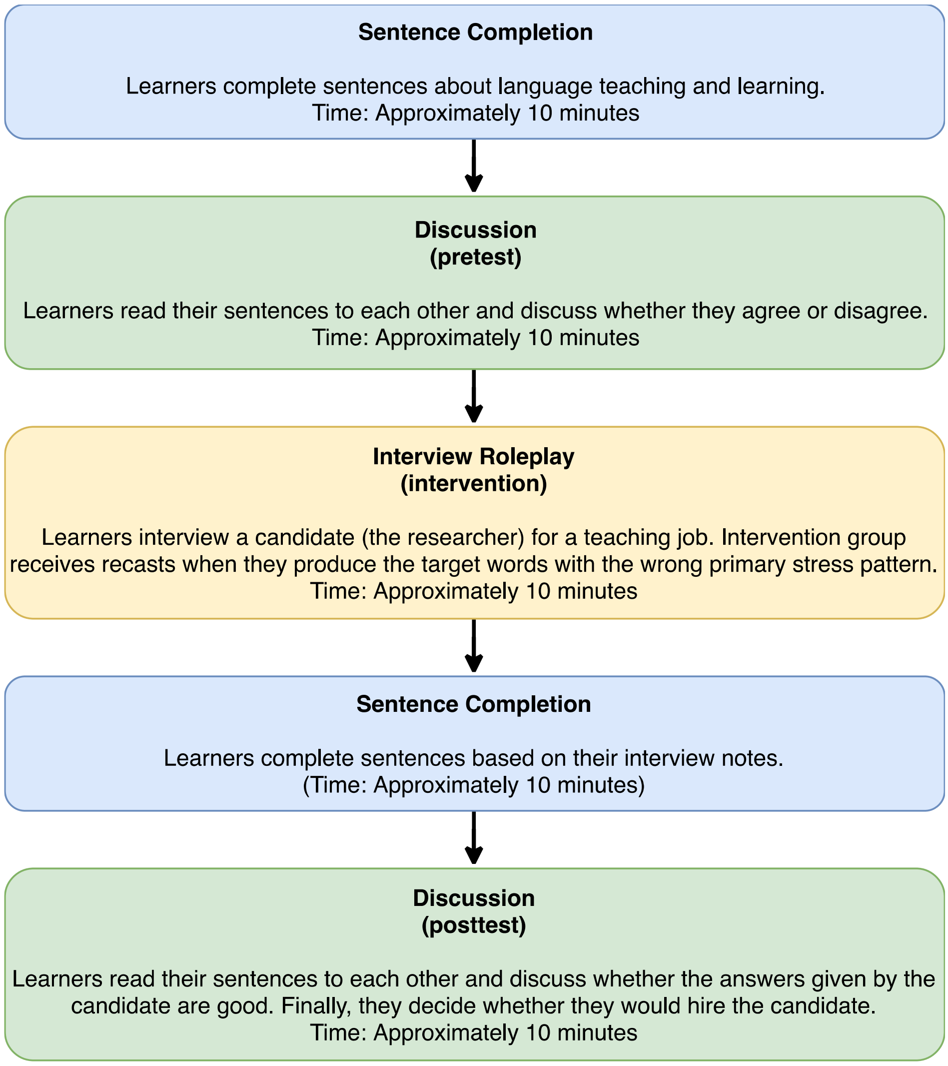


Figure 6.3 The procedure for data collection from learners

explanation without modeling the pronunciation of the target word. It is important to note that the researcher received only a few questions about word meaning whether it was about the target vocabulary or any other word that appeared in the handouts. Learners completed the handout on their own and they did not interact with their partner at this stage.

After learners finished the first sentence-completion activity, they carried out the information exchange task. They took turns to read their sentences aloud to each other and commented on whether they agreed or disagreed with their partner's opinion in the form of a brief discussion. Learners' productions of the target words during this interaction formed the pretest data.

The next activity was the interview task. Having discussed issues related to language teaching and language classrooms, learners played the role of a recruitment committee to interview the researcher, who pretended to have applied for a job at the Intensive English Program where learners were actually studying. Learners were given the handouts with the interview questions, which had the target words embedded in them. They took turns to ask their questions and they took brief notes while listening to the answers provided by the researcher. When learners in the intervention group produced a target word with a non-target-like stress pattern, the researcher provided a recast as explained in Chapter 6.5. Learners in the intervention group received a total of 199 recasts throughout the entire data collection period. Considering that the 36 learners produced a total of 360 tokens, they received recasts on 55% of the total number of target words that they produced.

After the interview task, learners were given the last handout, which was a sentence-completion activity that required learners to complete a set of sentences based on the responses given by the researcher during the interview. Learners completed the handouts on their own. These handouts were then used to facilitate the post-interview discussion.

Finally, learners shared their interview notes with each other by taking turns to read them aloud. During this task, learners discussed whether they liked the responses given by the researcher during the interview. At the end, they decided whether they should offer the researcher a job to teach at the Intensive English Program. Learners' production of the target words during this interaction formed the posttest data.

During the information exchange and interview tasks, learners wore a Shure WH20 XLR brand unidirectional microphone mounted on their head. As learners carried out the activities in pairs, two headworn microphones were used. The microphones were connected to a Roland Duo-Capture EX USB audio interface. The two microphones were connected to separate input channels, which allowed the speech data from each learner to be recorded separately. The microphones were positioned at about 30 degrees off-axis and about 3 cm away from learners' mouth similar to the procedure that was followed by Kirkham (2017). The gain levels were kept constant through the entire data collection period. The recordings were saved onto an iPad Pro as 44,100 Hz 16 bit .wav files using a recording app called Recorder Plus HD.

Upon completion of the second information exchange and consensus task, learners completed the exit survey on a laptop computer. Completing the exit survey took learners around 10 minutes. Once again, learners were given an opportunity to ask questions when they did not understand a survey item or a particular word. At the end of the data collection, the researcher thanked learners and presented them with the voucher for free coffee and donut. Each data collection session lasted about one hour. The entire data was collected in two consecutive semesters. The first and bigger portion of the data was collected from 60 learners over a period of 1.5 months in Fall 2016. The second portion of the data was collected from 14 learners over a period of 3 weeks in Spring 2016. Although the data was collected in two different semesters, both groups of learners had similar language proficiency levels. They were also registered in the same Intensive English Program level, but in two different semesters. In other words, Spring 2016 learners were not continuing students; they were new intake.

6.6.2. Data collection from listeners

The data from listeners were collected individually in each listener's office, which was quiet during data collection. As the first step, listeners were informed about the data collection procedure and then they were given the consent form to sign. After that, they completed a brief background survey on Qualtrics (2018) using a laptop computer provided by the researcher (see Appendix C). This was followed by the recording of speech samples that were later used for creating the native speaker baseline against which learner productions would be compared. Listeners were given 20 sentences that were previously written by learners as part of the second sentence-completion activity. The sentences were checked for grammar. Minor grammar issues were corrected by keeping alterations at minimum in order to maintain the original syntax. Listeners were asked to read aloud each sentence at a normal speed. Their speech data was recorded using the same equipment that was used for collecting the speech data from learners.

Once their speech data was recorded, listeners proceeded with the rating of 120 words produced by learners. The number of words selected for the listener judgment experiment was limited to 120 in order to minimize the potential negative impact of listener fatigue on the integrity of the ratings. There can be various reasons for listener fatigue, and in the current study, a potential trigger for listener fatigue was increased listener effort. Listener effort refers to the attentional requirements needed to understand speech (Fraser, Gagné, Alepins, & Dubois, 2010; Hicks & Tharpe, 2002; McGarrigle et al., 2014) and it may be adversely affected by accented speech (Mattys, Davis, Bradlow, & Scott, 2012) and a long rating session (Saito & Lyster, 2012a). Although, the SLA literature does not provide any duration-specific guidelines for conducting listener ratings, Mackey and Gass (2016) suggest that it is possible to achieve meaningful interrater reliability results with 10% of the data. Therefore, 120 words were deemed appropriate for practical and statistical purposes. The 120 words

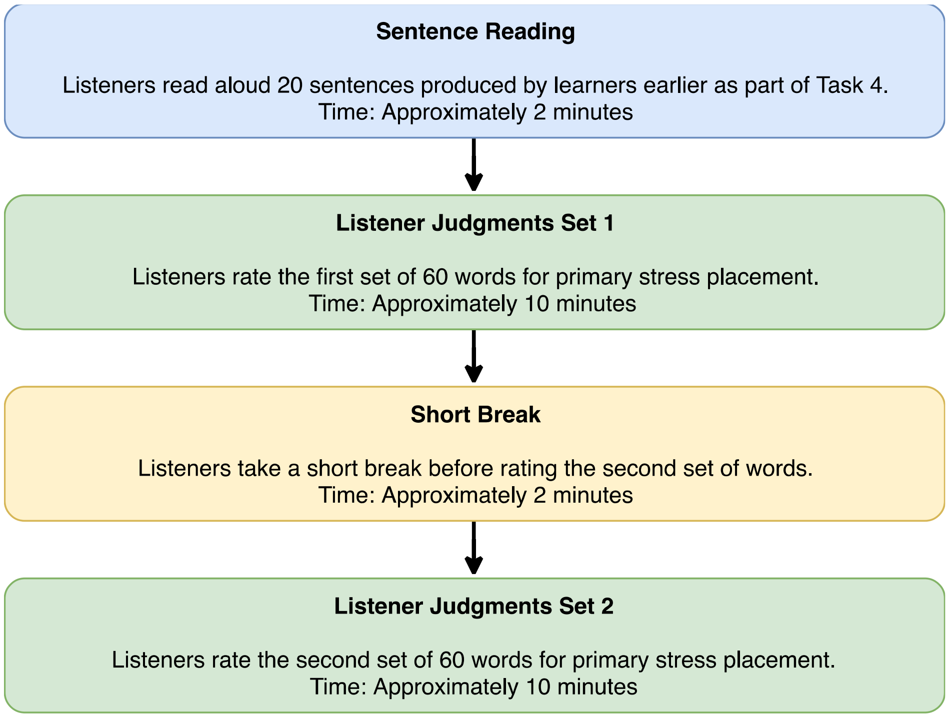


Figure 6.4 The procedure for data collection from listeners

comprised 30 pretest-posttest pairs of words produced by the control group and 30 pretest-posttest pairs of words produced by the intervention group. The word pairs selected from the intervention group had all received a recast during the intervention. In preparation for the rating sessions, 10 different randomizations of the 120 words were created for the 10 native-speaker listeners. Then, each randomization was divided into two equal sets to allow listeners to take a short break half way through the rating session. Prior to making judgments of stress placement, listeners completed a short training session on PsychoPy (Peirce, 2009) focusing on 6 words. The main purpose of the training session was to allow listeners to familiarize themselves with the software and the procedure for judging stress placement in order to reduce possible rater error attributable to task novelty. As part of the training, listeners were also provided with brief background information on primary stress in English. The 6 words that were used for the training session were not used for the actual experiment. Figure 6.4 shows the procedure followed when collecting data from listeners.

When listeners were ready to proceed to the listening experiment, they wore JBL Everest Elite 700 model noise-cancelling headphones and started rating the first set of 60 randomized words. When they were done, they took a short break of 2-3 minutes and then rated the second set of 60 randomized words. Listeners were allowed to replay and listen to a word as many times as they wanted by pressing the R key on the keyboard. While they were listening to a word, the nominal scale used for the judgment of stress placement was displayed on the screen. When listeners completed rating a word for stress placement, they pressed the N key on the keyboard to proceed to the next word. The researcher was present during the entire rating session for each listener to ensure consistency in the rating procedure and to provide support in case listeners faced a technical problem. All 120 words were rated in one session, which took approximately 20-25 minutes. All listeners used the same Macbook Pro computer provided by the researcher when carrying out the rating experiment and no technical problems were faced. The responses from each session were automatically saved into a .csv file.

6.7. Acoustic analysis

The first set of analyses focused on duration, mean intensity, and mean pitch values extracted from the syllables.

6.7.1. Coding of acoustic measures

The recordings on the iPad were transferred onto a MacBook Pro laptop for coding and analysis. After the removal of data from 1 Hindu and 5 Arabic speakers due to the reasons discussed in Chapter 6.3.1, there were 36 participants in the intervention group and 32 participants in the control group. Using Audacity (Audacity Team, 2018), the two-channel recordings from each data collection session were split into two separate mono tracks. Each mono track contained the recording from one learner. The individual tracks were then loaded onto Praat (Boersma & Weenink, 2018) to display their waveform and spectrographic view for acoustic analysis. The researcher listened to each speech stream to identify the target words and then used the syllable boundaries displayed in Table 6.4 as a guideline when coding the data.

Table 6.4

Syllable boundaries of the target words

Set 1	dɑɪ.næm.ɪk	dɪ.vɛl.əp	ɹɪ.vɪʒ.ən	kən.sɪd.ɚ	səs.pɪʃ.ən
	kən.fju.ʒən	pər.sɛp.ʃən	ɹɪs.pɑn.sɪv	kən.tɛn.ʃən	kən.sɪs.tənt
Set 2	pə.zɪʃ.ən	dʊ.reɪ.ʃən	də.mɪn.ɪʃ	fɔɹ.meɪ.ʃən	kən.dɪʃ.ən
	kəmˈpæʃ·ən	kəm.pæn.jən	kən.sɛn.səs	kən.vɛn.ʃən	pər.sɪs.tənt

Some researchers have chosen to analyze the duration of vowels to determine changes in syllable duration (e.g., Chen et al., 2001; Gordon-Salant, Yeni-Komshian, Pickett, & Fitzgibbons, 2016). The current study, however, did not break syllables into vowels and consonants when analyzing any of the three acoustic measures. One of the reasons for this decision is that preceding and following consonants can have a considerable impact on vowel production (Cutler, 2005; Ladefoged, 2006), and in the same way, vowels can have an impact on the production of consonants. Measurements based on syllables rather than segments have also been used by Winter and O'Brien (2013) who explain that "increased prominence of accentuation applies to whole syllables, rather than individual segments" (p. 488). Therefore, it was deemed more appropriate in the context of the current study to treat each syllable as a complete unit. Needless to say, there were instances of individual differences among learners in terms of their production of particular segments, such as issues with voicing (e.g., producing /b/ instead of /p/) or being non-rhotic (e.g., producing the word persistent as /pə.sɪs.tənt/). The analysis did not delve further into these individual differences and mainly focused at the changes that occurred at the syllable level. That said, it is worth mentioning that the individual differences were in fact handled by the linear mixed-effects analyses that were used for the current study, which had learners entered into the model as random effects with random slopes of learner-by-time. Further details on statistical analyses are provided in the next section. When identifying syllables for the analysis, the syllable boundaries were mainly marked according to the maximal onset principle (Deterding, 2001; Duanmu, 2008) with the exception of target words that contained word-medial syllables with a short vowel. Some scholars such as Roach (2000) and Taft (2002) point out that English phonotactics does not allow short vowels (e.g., /æ/, /ɛ/, /ɪ/) to occur at the end of a syllable. Therefore, the maximal onset principle was partially followed when marking the syllable boundaries for the target words *condition, consider, compassion, develop, diminish, dynamic, position, revision,* and *suspicion*.[2] The target words and their syllable boundaries were marked using visual and auditory information. The researcher listened to each target word multiple times to ensure that word boundaries did not include any preceding or following sounds or pauses. The word initial boundary was marked at the onset of the stop release for target words starting with the voiceless stops /p/ and /k/; at the onset of voicing for the voiced stop /d/ and the approximant /ɹ/; and at the onset of frication for the voiceless fricatives /f/ and /s/. The word final boundary was marked at the offset of phonation for target words ending with the nasal /n/ and the approximant /ɹ/; at the offset of frication for the fricatives /ʃ/, /v/, and /s/; and at the offset of the release burst for the voiceless stops /p/, /t/, and /k/. Syllable boundaries were marked

[2] Before I had the opportunity to read Roach (2000) and Taft (2002), an earlier version of the coded data had fully complied with the maximal onset principle. The first set of analyses had been run based on this data set. However, after having read about phonotactic constraints regarding short vowels, the syllables that had a short vowel were recoded and the analyses were rerun. The comparison of the first and the second sets of analyses showed that there were no differences in group or individual patterns. Both sets of data yielded the same statistical results.

following the same principles. When coding target words such as *companion* and *consistent,* which have a penultimate or final syllable with a stop consonant in the onset position that creates a stop gap between the penultimate or final syllable and the preceding syllable, the word-medial stop gaps were coded as part of the penultimate or final syllable that carried the stop consonant. Figure 6.5 provides an example of how the word *companion* was coded.

The annotations that contained markings of syllable boundaries were saved as a TextGrid file. When there was a problem with a word, the problem was annotated in Tier 3 of the TextGrid file, which also included the orthographic form of the word. The sound file for each target word along with the attached annotations were extracted from the original recording and saved as an individual .wav file using a Praat script written by Kirkham (2014). Some of the problems encountered during the coding phase were production of a different word (e.g., “conversation” instead of “convention”), unintelligible production (e.g., [kənsɪnɪəns] for “consensus”), one phoneme change (e.g., [bəzɪʃən] instead of /pəzɪʃən/), missing phoneme (e.g., [kənsɪstən] for “consistent”), epenthesis (e.g., [səspɪkʃən] for “suspicion”), and syllable lengthening due to elongation of fricative or nasal consonants. Intelligible tokens that differed from the target word by a single phoneme were kept in the dataset. The different words as well as unintelligible or non-target-like productions were removed from the data set along with their corresponding pretest or posttest production, even if their corresponding pretest/posttest production was produced correctly or intelligibly. After the data cleansing process, there were a total of 1030 tokens; 261 pretest-posttest pairs produced by the control group and 304 pretest-posttest pairs produced by the intervention group.

Next, using another Praat script written by Kirkham (2015), the duration, mean

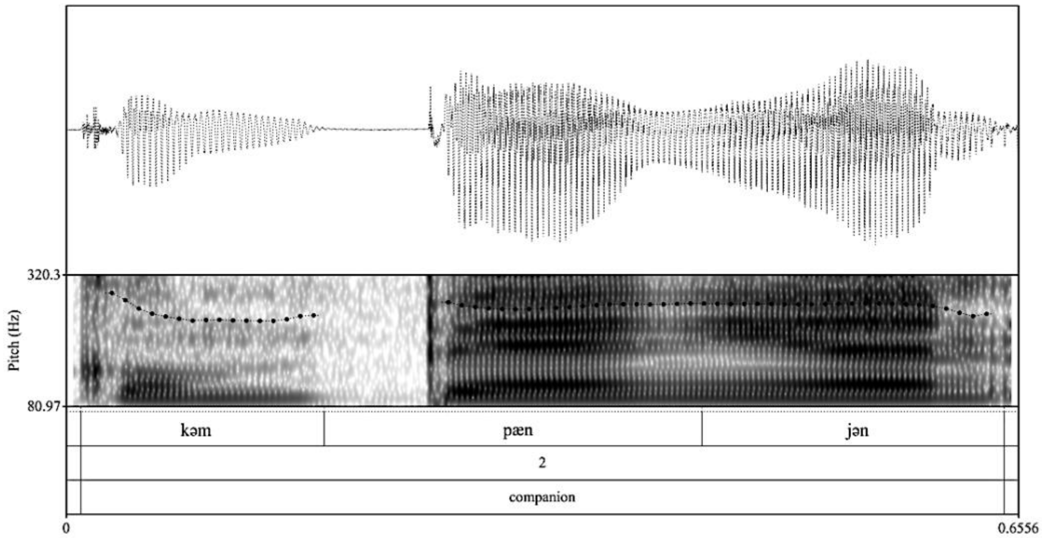


Figure 6.5 Marking of syllable boundaries

intensity, and mean pitch measurements for each syllable were extracted from individual words and saved as a .csv file. The rationale for focusing on duration, intensity, and pitch was that research has confirmed that these acoustic cues are reliable correlates of lexical stress in English as discussed in Chapter 4.3. The duration measurements, which were in milliseconds, were multiplied by 1000 and converted to seconds. The mean intensity measurements extracted in decibels (dB) were kept in their original form. On the other hand, the extraction of the mean pitch measurements required the implementation of a few additional steps.

Pitch extraction is not always a straight-forward process as the algorithm used for pitch extraction can suffer from the smallest problems in the sound signal (Gerhard, 2003; Zsiga, 2013). The data collected for the current study did not present any problems in this regard because the technical features of the recording equipment and the quiet recording environment ensured problem-free recordings. However, to preempt any unforeseeable issues with pitch extraction, additional steps were taken. Following the parameter optimization method recommended by Keelan, Lai, and

Zechner (2010), individual pitch floor and ceiling were determined for each speaker to increase the accuracy of pitch measurements. To begin with, a speech sample of approximately 30 seconds including minimal number of pauses was selected for each learner. Then, the pitch distribution across the selected speech sample was extracted using the Pitch Listing function in Praat. Prior to the extraction, the pitch floor was set to 75Hz and the pitch ceiling to 600Hz. After that the 35^{th} and the 65^{th} quantiles were determined in order to calculate individual pitch floor and pitch ceiling values following the formulas below (as cited in Keelan, 2010):

Pitch Floor = q35 x 0.72 – 10

Pitch Ceiling = q65 x 1.9 + 10

The individual pitch floor and pitch ceiling values for each learner were entered into the settings when extracting the mean pitch measurement from syllables. The analysis of pitch also needs to account for biological differences between men and women that affect the pitch ranges for each gender. Adult men have longer and thicker vocal folds which vibrate at a slower rate compared to the vocal folds of adult women (Zsiga, 2015). As a result, the voice of an adult male may have a pitch range of 80-200Hz with an average of 120Hz, whereas the voice of an adult female may cover a range of 180-400Hz with an average of 220Hz (Couper-Kuhlen, 1996; Simpson, 2009). In addition to the biological differences that affect male and female pitch ranges, there is a mismatch between the linear representation in hertz and logarithmic perception of pitch by humans that needed to be considered. The value of pitch, or fundamental frequency in acoustic terms, is based on the vibrations of the vocal folds and it is represented in hertz. As calculation of pitch is directly tied to vocal fold vibration cycles per second, pitch represented in hertz is a linear measure. However, the way the human ear perceives changes in pitch levels is not linear. Humans perceive changes in pitch logarithmically (Couper-Kuhlen, 1996). In other words, a speech sample produced at 200Hz is not perceived by the human ear twice as high pitched compared to a speech sample produced at 100Hz. In order to account for this difference and for meaningful comparisons from a perceptual perspective, intervals need to become wider at higher frequencies. This necessitates the conversion of pitch values in hertz to semitones (Couper-Kuhlen, 1996; Simpson, 2009). Considering these points, the pitch measurements in Hertz were converted to semitones using the f2st function in R using the syntax below, which is based on the h2st function developed by Mark Lieberman (Quené, 2015):

f2st(vector containing values in hertz)

The final step of data preparation was normalizing the data by calculating difference ratios. As it has been mentioned earlier in the literature review, perception and production of lexical stress needs to be discussed in terms of the relationship between the stressed and unstressed syllables in the same word. Fry (1958) explains that this type of relativity is expected as perceived stress is unaffected by the changes in speech rate; that is to say, the same word can be produced slow or fast, or in a quite manner or in a loud manner by the same speaker and none of these changes will affect the perceived prominence of primary stress. Therefore, it is important to consider changes in stress from a ratio perspective. Measures based on a ratio calculation are also a method of normalization and they help control for the variations that could occur due to differences in speech rate between two speakers. For example, a speaker who speaks at a fast rate will produce a particular word in a shorter period of time compared to a speaker with a slower speech rate. As speech rate has a direct impact on syllable-level duration, the stressed syllable produced by the fast speaker will be shorter than the stressed syllable produced by the slow speaker; however, that does not mean that the fast speaker produces the stressed syllable with less stress. Therefore, any analysis that has an inter-speaker element needs to account for differences in speech rate. One last utility of ratio-based measures is that they allow for comparisons between different types of acoustic correlates. In the current study, duration was originally measured in milliseconds, pitch in Hertz, which was later converted to semitones, and intensity in decibels. As it would not be possible to compare the changes in one acoustic measure to another when they are represented on different scales, it was necessary to use a ratio measure to be able to compare the amount change that occurred in one acoustic correlate to the amount of change that occurred in another acoustic correlate.

In order to account for the acoustic and perceptual relativity as well as to normalize the data, it is possible to choose from three types of ratios measures. The first measure is a simple difference measure obtained by subtracting unstressed syllable(s) from the stressed syllable. The second measure is a ratio measure obtained by dividing the stressed syllable by the unstressed syllable(s) (e.g., Flege & Bohn, 1989; Fry, 1955, 1958; Nguyễn & Ingram, 2005; Zuraiq & Sereno, 2007). Finally, the third measure is a relative difference measure calculated by subtracting unstressed syllable(s) from the stressed syllable and then dividing the difference by the sum of stressed and unstressed syllables (e.g., Gordon-Salant et al., 2016). Taylor and Wales (1987) labelled these three types of ratios as subtraction ratio, division ratio, and Michelson contrast ratio respectively. They argued that the contrast ratio, which had been originally introduced in the field of optics as a measure of visual contrast, is a more robust predictor of accent compared to subtraction and division ratios. Their study attempted to identify the acoustic features that predict perception of accent, which was placed on either a modal verb or a main verb. Based on previous literature, they chose seven acoustic features to include in the analyses, namely, duration, F_0 peak, F_0 mean, F_0 peak minus mean, intensity peak, intensity mean, and intensity peak minus mean. They ran three analyses using the three ratios. There were two important findings in their study. First, the amount of covariation between different predictors changed depending on the ratio that was used. The subtraction ratio resulted in the highest amount of covariation compared to the division ratio and the contrast ratio. Second, contrast-based acoustic correlates were more successful at predicting accent placement for both modal verbs and main verbs. Therefore, the contrast ratio was deemed suitable for the current study due to the fact that the focus was a form of prosodic prominence. Taylor and Wales (1987) explored the realization of accent and analyzed duration, pitch, and intensity as predictors. In their study, accent fell on a single syllable, as it normally does, which has parallelisms with the current study as accent is placed on the syllable that carries primary stress.

In the current study, the normalization was carried out using the contrast ratio, which is referred to as *proportional difference measure*. In order to calculate the proportional difference measure, the difference between Syllable 2 (S2) and the sum of Syllable 1 (S1) and Syllable 3 (S3) was divided by the sum of all three syllables. The calculation was carried out separately for each acoustic correlate. Below is the formula for the calculation:

$$S_{norm} = (S2-(S1+S3))/(S1+S2+S3)$$

The rationale behind this calculation was to capture the acoustic changes that may have occurred across all three syllables rather than focusing only on the changes that occurred in S2, which could have undermined the integrity of the analysis.

6.7.2. Statistical analysis of acoustic measures

For statistical analyses of the acoustic measures, a linear mixed-effects model with restricted maximum likelihood was fitted with *time* and *condition* as fixed effects, and *learners* and *words* as random effects using the lme4 package version 1.1-14 (Bates et al., 2015) in R Studio (Rstudio Team, 2016) working on R version 3.3.3 (R Core Team, 2017). Restricted maximum likelihood estimation is a recommended method for mixed-effects models as it provides estimates with a higher level of precision (Baayen, Davidson, & Bates, 2008). The mixed-effects models included random intercepts as well as random slopes. A more detailed discussion on how the models were fitted for the analysis of duration, intensity, and pitch measures is provided in the next section (Chapter 6.7.3). In this section, the rationale behind the choice of mixed-effects analysis is explained from a methodological perspective. Also, the assumptions that were checked prior to conducting any statistical analysis as well as the assumptions that needed to be checked following the analyses (e.g., distribution of residuals) are discussed. Before getting into the details on the choice of statistical tools and the motivations behind it, it would probably be useful to provide a brief definition for each of the terms that will be used in this section as well as the upcoming sections.

The theory of mixed-effects models is based on three elements: fixed effects, random effects, and residuals. Residuals are technically a part of random effects as it will be explained later on. The representational formula below shows the role of these three elements in determining the outcome variable based on the theory of mixed effects models (for a more technical discussion see West, Welch, & Galecki, 2015).

outcome variable = fixed effects + random effects + residuals

If the terms used in the given formula were expressed in terms that are more common in the SLA field, the outcome variable would be the dependent variable (also known as the response variable) and fixed effects as well as random effects would be the independent variables (also known as explanatory variables). The dependent variables are the outcome of an experiment, in other words, the data that has been collected; and independent variables are the variables that influence that outcome (Larson-Hall, 2016). Researchers may manipulate or control for independent variables to understand their predictive value on the dependent variable. For example, in the current study, the three acoustic correlates of primary stress are dependent variables. On the other hand, membership to one of the two experimental conditions is an independent variable as learners in the intervention group received recasts on their primary stress errors and learners in the control group did not. In the data frame and in the syntax used for the mixed-effects analyses, this independent variable was represented as *condition.* Furthermore, learners in both groups carried out two sets of tasks at two different points in time from which pretest and posttest data was collected. The first time was prior to the intervention and the second time was after the intervention. These two data collection times are considered as another independent variable and were represented in the data frame and model syntax as *time.* When carrying out the acoustic analyses, condition and time were treated as fixed effects. It is important to note that fixed effects are the effects that do not vary and the levels of a fixed effect represent specific conditions. In the current study, the two fixed effects had two levels each. The levels for *condition* were *intervention* and *control,* and the levels for *time* were *pretest* and *posttest.* As can be understood from these examples, the goal of a study is not to generalize over the fixed effects. Random effects, on the other hand, are the effects that the researcher wants to "generalize beyond the parameters that constitute the variable" (Larson-Hall, 2016, p. 481). In the current study, the random effects were *learner* and *word* because they were randomly selected from the populations that they represent. The goal of the study was to generalize the effects of recasts measured by the interaction between condition and time beyond the learners that took part in the current study as well as beyond the three-syllable words that were used in the current study. Finally, residuals are the amount of variance that is not explained by fixed effects or random effects, in other words the model itself. When creating the syntax for a mixed-effects analysis, fixed effects are entered into the first part of the syntax and random effects are entered into the second part. Technically, residuals cannot be entered into the syntax because they are an outcome of the analysis.

The choice of linear mixed-effects analysis over t-test analysis was motivated by several considerations. The first set of considerations were mainly related to the nature of the data and the assumptions for carrying out appropriate statistical analyses. The data set contained more than one data point, that is multiple target words, from each participant on both the pretest and posttest. Also, each target word was produced by learners on both the pretest and the posttest. This situation is a violation of the independency assumption, an assumption central to t-test and ANOVA analyses, which requires each data point to be independent from the rest of the data points. Also, participants were recruited from several classes at the Intensive English Program and one could argue that two participants who studied in the same classroom are not independent. This could also be considered as a violation of the independency assumption. Moreover, the observations were not equally distributed; to be more specific, some participants produced their complete set of 10 target words on both the pretest and the posttest, whereas others either skipped a few words or in some cases target words were removed due to intelligibility problems. As a result, the target words were not evenly distributed across participants and groups creating an unbalanced data set, which is also a violation that prevents a t-test analysis. Although learners' productions on the pretest and the posttest could have been averaged resulting in two values for each learner, this would still be problematic as the average value for some learners would have been based on, say, 6 words whereas for others that would be 8 or 10 words. Because of these issues, it was not possible to use a t-test to analyze the data. On the other hand, as a more advanced statistical method, linear mixed-effects analysis can handle non-independencies and unbalanced data with the help of random effects parameters (Linck & Cunnings, 2015; Quené & van den Bergh, 2008; Winter, 2013). Linear mixed-effects models are capable of creating by-subject or by-item intercepts when these variables are entered into the model as random effects. This allows for the handling of the multiple data points associated with a subject or an item together in order to model the random variation in by-subject or by-item averages. For the current study, mixed-effects modeling made it possible to treat learners and target words as a random effect allowing the fitted model to make by-learner and by-word adjustments to the intercept, which forms the baseline mean in the case of linear mixed-effects models. Another consideration during the model fitting phase was the inclusion of random slopes. The current study follows a pretest-posttest design, as such, the target words were produced by learners twice; thus, the study has a repeated-measures design. In this context, the variance among learners or words on the pretest may be different from the variance on the posttest. For example, learners who have a similar pretest performance may differ to a greater extent on the posttest as some learners may benefit more from recasts and display higher gains on the posttest whereas others may benefit to a lesser extent and display lower gains.

Likewise, words that are produced with a similar S2 duration on the pretest may differ in terms of S2 duration on the posttest. Therefore, considering potential by-learner and by-word heteroskedasticity, the model needed to be adjusted to account for the differences in the effects of time on individual learners and individual target words. Once again, this is something that a mixed-effects analysis can handle. In order to account for by-learner and by-word heteroskedasticity, learner-by-time and word-by-time random slopes were included in the models in addition to the random intercept. The inclusion of random slopes also prevents overconfident results and the possibility of a Type I error when fitting a linear mixed-effects model for a study with a repeated measures design (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013; Cunnings & Finlayson, 2015; Schielzeth & Forstmeier, 2009).

The second set of considerations that led to the choice of running a mixed-effects analysis were theoretically motivated. Entering words as a random effect into the model addresses the problem of language-as-fixed-effect fallacy that was brought to attention by Clark (1973) a couple of decades ago. Clark argued that when the set of language stimuli in a study is not treated as a random effect, it would be statistically unsound to generalize beyond that specific set of language stimuli, even if they were randomly selected. The current study focuses on a particular set of 20 words, and to be able to generalize beyond those 20 words and to make confident inferences, it was necessary to use a mixed-effects model and treat words as a random effect. In the same way, entering learners into the model as a random effect helps with making inferences that go beyond the sample of Arab learners who participated in the study. In summary, learners and words were treated as partially crossed random effects because half of the learners in the intervention and control groups produced one set of 10 words and the other half produced another set of 10 words. In this context, the term crossed means that the learners were tested on the set of words selected for the study, and at the same time, the set of words were tested on the learners, which directly addresses the problem of language-as-fixed-effect fallacy (for a detailed discussion see Baayen et al., 2008; Cunnings, 2012; Cunnings & Finlayson, 2015; Locker, Hoffman, & Bovaird, 2007, Winter & Wieling, 2016). It is worth mentioning that linear mixed-effects models have been used in a number of recent applied linguistics studies (e.g., Gordon & Darcy, 2016; Murakami, 2016; Rogers, 2017; Tremblay, Derwing, Libben, & Westbury, 2011) due to their flexibility and the advantages mentioned above. Following this trend, the choice of a mixed-effects analysis was also an attempt to contribute to the advancement of statistical methods used in quantitative L2 research, a need that has been highlighted recently (Plonsky, 2013, 2015; Plonsky & Gass, 2011; Plonsky & Oswald, 2016).

Once the method of inferential statistics has been chosen, a crucial step prior to analysis is assessing the assumptions. Therefore, the assumptions of normality (both for raw data and model residuals), linearity, and homoscedasticity were checked as part of the analysis. Normality is probably the least important of all the assumptions when running a linear mixed-effects analysis as the method itself is robust for violations of normality. In fact, some statisticians (e.g., Gelman & Hill, 2006) do not recommend checking normality when running regression analyses and others ignore checking distribution of data altogether and mainly focus on the distribution of residuals for diagnostic purposes (e.g., West et al., 2015). Yet, in order to maintain transparency in reporting of the procedures and analyses used in the current study, normality was checked using q-q plots prior to running the analyses. After fitting the models, the distributions of residuals were also checked. As recommended by Zuur, Ieno, and Elphick (2010), visual exploration of the data was carried out instead of running any tests of normality. The distributions of duration and intensity measures were fairly close to normal distribution. The distributions for pitch measures were slightly tailed; however, this was not seen as a problem as linear mixed-models are robust against violations of normality (Winter, 2013). The distribution of the data as well as the distribution of the residuals for each model can be seen in Appendix E.

Finally, the assumptions for linearity and homoscedasticity were checked using residual plots. The more important of the two assumptions is linearity as linear mixed-effects models are also robust against violations of homoscedasticity, an assumption central to linear regression models (Linck & Cunnings, 2015; Quené & van den Bergh, 2008). The residual plots did not indicate a violation of either of the assumptions (see Appendix G).

6.7.3. Model fitting for acoustic analysis 1

The first set of acoustic analyses focused on the pretest-posttest differences in the performance of the intervention and the control groups. A series of linear mixed-effects models were fitted for the analysis of duration, intensity, and pitch measures. A recommended method for fitting a linear mixed-effects model in the case of a confirmatory research study is to create the maximal model that takes the study design and the research questions into consideration (Barr, 2013; Barr et al., 2013; Cunnings & Finlayson, 2015). Following this recommendation, the formula below was used when fitting the maximal model for the analysis of the proportional difference in duration:

maximal model for duration: lmer (duration ~ condition * time + (1 + time | learner) + (1 + time | word))

In this formula, the outcome variable *duration* appears on the left side of the tilde operator (~), and on the right side of the tilde are *condition* and *time,* which are the fixed effects. The star sign between condition and time allows the model to account for the interaction between two fixed effects. It should be noted that condition and time are crossed effects as both experimental conditions have data points that correspond to the two levels under time, which are pretest and posttest. Because the main goal of the study was to observe changes in acoustic measures over time, the interaction between time and condition as fixed effects was the main interest of the analysis. The second part of the formula shows that the variables *learner* and *word* are entered into the model as a random effect with their own intercept. The two random effects are partially crossed as half of the learners in both experimental conditions produced one half of the target words and the other half in both conditions produced the other half of the target words. As can be seen in the formula, time was also used in the second part along with the two random effects. This allows the model to create the random slopes of learner-by-time and word-by-time as it was discussed earlier in Chapter 6.7.2.

However, it was not possible to use the maximal model for the analysis of syllable duration because the model failed to converge; and as a result, adjustments needed to be made. Convergence issues are not uncommon when fitting mixed-effects models and they usually occur due to model complexity. Model complexity is a relative term and it is directly related to the amount of data used for model estimation. When the model is too complex with a number of random intercepts and random slopes, the model will not be able to estimate the correlations between the random intercepts and the random slopes as it will run out of degrees of freedom. It is possible to handle this issue by simplifying the model. Among the various possibilities for model simplification, there are two common methods. One method is identifying and removing the random effect that has the lowest variance. This method works well in the case of explanatory data analysis, a case where researchers try to discover patterns in the data, sometimes using a method of trial and error. However, in the current situation, the mixed-effects model is based on a particular research design and it serves the purpose of confirmatory analysis. Since both random effects are crucial in terms of generalizing findings and the issue of language as a fixed-effect fallacy as it was discussed in Chapter 6.7.2, removing the random effect learner or word was kept as a last resort. As an alternative option, Cunnings and Finlayson (2015) recommend simplifying the model by removing the correlation between a random intercept and a random slope. They explain that when the correlation parameter is present, the model tries to account for the possibility that a higher intercept may also have a higher slope. To give an example, a learner who produces stressed syllables with a higher average duration than other learners may exhibit higher gains in duration on the posttest. In the same way, a word that has been produced with a longer second syllable on average may display higher gains in duration on the posttest. In the current model, the syntax *1 + time* generates a random intercept and a random slope as well as the interaction between the two. As there were two random effects in the model, and hence two correlations, as a first step the model was simplified by removing the correlation parameter from the random effect *word* using the following expression: (1|word) + (0 + time | word). The complete version of second model can be seen below.

> second model for duration: lmer (duration ~ condition * time + (1 + time | learner) + (1|word) + (0 + time | word)

However, the second model also failed to converge, and therefore, the model was simplified further by removing the correlation parameter from the random effect *learner*. The third model converged and no error messages were received. Thus, the third model, which is provided below, was used for the analysis of duration measures.

third model for duration: lmer (duration ~ condition * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

For the analysis of syllable intensity, once again the maximal model was fitted similar to the process that was followed for the analysis of syllable duration:

maximal model for intensity: lmer (intensity ~ condition * time + (1 + time | learner) + (1 + time | word))

The maximal model fitted for the analysis of intensity failed to converge, necessitating the use of the model simplification methods that were followed for the analysis of duration. After removing the correlation parameter from the random effect *word*, the model still failed the converge. Thus, the model was simplified further by removing the correlation parameter from the random effect *learner*. The third model converged successfully and was used for the analysis of intensity (see Appendix F).

third model for intensity: lmer (intensity ~ condition * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

The third and last part of the first set of analyses focused on the pitch measures. Once again, the maximal model was fitted for the linear mixed-effects analysis of pitch measures. Unlike duration and intensity measures, the maximal modal converged without any errors in the case of pitch measure (see Appendix F).

maximal model for pitch: lmer (pitch ~ condition * time + (1 + time | learner) + (1 + time | word))

6.7.4. Model fitting for acoustic analysis 2

The second set of acoustic analyses focused on the intervention group in order to compare the acoustic realization of the target words that received a recast and the target words that did not. There were two reasons behind the need for running a separate analysis to examine the impact of feedback as a fixed effect. The first reason was the conceptual differences between the target words that did not receive a recast. More specifically, the target words that did not receive a recast were conceptually different from one another depending on which group of learners they were produced by. As the learners in the control group did not receive any form of feedback during the interview activity, every single target word they produced was labelled as *no recast* when coding the *feedback* variable, regardless of whether the word exhibited more target-like lexical stress placement or not. When it comes to the intervention group, however, the *no recast* label was only applied to the words that were produced with more target-like stress as those words did not receive a recast from the researcher. In other words, the reasons for withholding feedback were different depending on the experimental condition, and as a result, not receiving a recast on a target word was a different concept in the two experimental conditions. Furthermore, in the case of the intervention group, there could have been a priming effect on the target words that did not receive a recast as the learners who produced those words actually received recasts on other words that they produced. Last but not least, learners were also exposed to the recasts that their activity partner received.

In addition to the issue of conceptual differences, R would have given an error if *feedback* had been entered into the mixed-effects models as a fixed effect along with *condition* as a fixed effect. Feedback was treated as a categorical variable with two levels, namely, *recast* and *no recast.* That said, the two levels only corresponded with the target words produced by the intervention group, while there was only one level, *no recast,* for the words produced by the control group. Due to the structural differences in the feedback variable, entering both *condition* and *feedback* into the model as a fixed effect would have led to a rank deficiency error. Therefore, considering both the conceptual and the technical issues, a subset of the data comprising data points only from the intervention group was created for the second set of acoustic analyses.

When fitting the mixed-effects models for the second set of analyses, the same steps that were followed for the first set of analyses were followed. First, the maximal model was fitted for the analysis of duration using the formula below.

maximal model for duration: lmer (duration ~ feedback * time + (1 + time | learner) + (1 + time | word))

As the data points were only from the intervention group, instead of *condition*, *feedback* was entered into the model as a fixed effect. Similar to the between-groups analysis of the duration variable, the maximal model failed to converge. Thus, the model was simplified by removing the correlation parameter between the intercept and the slope of the random effect *word.* The second model also failed to converge. The third model, which was fitted without the correlation parameters for both random effects, converged and was used for the analysis of duration (see Appendix F).

third model for duration: lmer (duration ~ feedback * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

For the analysis of syllable intensity, once again the maximal model was fitted similar to the process that was followed for the analysis of syllable duration:

maximal model for intensity: lmer (intensity ~ feedback * time + (1 + time | learner) + (1 + time | word))

The maximal model fitted for the analysis of intensity failed to converge; therefore, the same steps for model simplification that were followed for the analysis of duration were followed in this case as well. After removing the correlation parameter from the random effect *word*, the model still failed the converge. Therefore, the model was simplified further by removing the correlation parameter from the random effect *learner*. The third model converged successfully and was used for the analysis of intensity (see Appendix F).

third model for intensity: lmer (intensity ~ feedback * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

Finally, for the analysis of pitch measures, once again the maximal model was fitted as the first step. The model converged and was used for the analysis of pitch measures (see Appendix F).

maximal model for pitch: lmer (pitch ~ feedback * time + (1 + time | learner) + (1 + time | word))

6.7.5. Calculation of goodness-of-fit measures

The R^2 value was calculated for each of the models fitted to provide a summary of the amount of variance explained by the model. Before explaining the method that was used to obtain the R^2 values, it is imperative to note that there is no commonly accepted way of obtaining an effect size measure for linear mixed-effects model. This is mainly due to the complexity of the mixed-effects analysis. Unlike a t-test analysis, linear mixed-effects models are much more flexible as well as complex as they can account for a number of interactions and generate model estimations based on random effects with their own intercepts and slopes. That is to say, the wealth of options and flexibility offered by mixed-effects models result in highly complex estimation analyses (LaHuis, Hartman, Hakoyama, & Clark, 2014). This situation makes the calculation of an effect size difficult. For example, it would be theoretically unsound to calculate Cohen's *d* for a mixed-effects model as the calculation of Cohen's *d* does not factor the variance explained by the random effects. There are also concerns about reporting R^2 as a goodness-of-fit statistic or as an effect size measure due to the multiplicity of the ways in which it can be calculated. Bates, the lead author of the lme4 package that was used in the current study for running the mixed-effects analyses, is critical of the attempts for defining an R^2 measure for linear-mixed models. In his response to a question on an online forum, Bates (2010) explains that the R^2 measure is for linear models that do not have random effects, and therefore, function differently from linear-mixed effects models. In practice, though, some linguistics or applied linguistics studies report R^2 values (e.g., Zellers, 2017) and others do not (Murakami, 2016; Rogers, 2017; Tremblay et al., 2011).

Recently, a group of scholars working in the field of Ecology and Evolutionary Genetics have been able to achieve some progress in calculating R^2 for mixed-effects models. Nakagawa and Schielzeth (2013) proposed two methods for obtaining R^2 for both linear and generalized linear mixed-effects models. The two methods they proposed were marginal R^2, which addresses variance explained by the fixed effects, and conditional R^2, which addresses variance explained by both fixed and random effects, in other words, the complete model. However, the methods they proposed did not cover an entire range of mixed-effects models because their formulae included random intercepts but not random slopes. Johnson (2014) expanded on Nakagawa and Schielzeth's work and updated the formulae so that they include the calculations for random slopes, which broadened the potential use of marginal and conditional R^2. This body of work culminated in the R package *piecewiseSEM* (version 2.0.1) created by Lefcheck (2016).

When the piecewiseSEM package was tested on the models created for the current study, it successfully processed the maximal model created for the analysis of pitch values. However, the package failed to process the models that were used for the analysis of duration and intensity probably because the models were too complex due to the syntax used to remove correlations from random effects. In order to tackle the issue, a simplified version of the duration and intensity models with just the random intercepts were entered into the syntax to obtain the R^2 values as can be seen below. The term *correlate* was replaced with duration and intensity for the respective analyses. The syntax below successfully produced the R^2 values for the duration and intensity models.

```
rsquared (lme4 :: lmer (correlate ~ time * condition + (1 | learner) + (1 | word))
```

However, since there were differences between the original lmer syntax used to obtain the model estimates and the rsquared syntax used to obtain a goodness-of-fit measure, an alternative method was also tested. The function below, also known as *r2.corr.mer,* was originally written by Byrnes (2008) and it provides a value for goodness of fit based on the correlation between the fitted and the observed values:

```
function(m) {
        lmfit <- lm(model.response(model.frame(m)) ~ fitted(m))
         summary(lmfit)$r.squared
}
```

As can be seen in the syntax above, this function is also based on linear models, which means its usage will be susceptible to the criticism raised by Bates (2010). Keeping the caveats in mind, the r2.corr.mer function was used to obtain R^2 values for duration, intensity, and pitch models with the intention of comparing those values to the ones that had been obtained through the piecewiseSEM package. The function successfully worked on all three models that had been originally fitted for the analysis. Table 6.5 provides a comparison between the R^2 values obtained by using the piecewiseSEM package to the ones obtained by the r2.corr.mer function.

Table 6.5

Comparison of R^2 values obtained for Acoustic Analysis 1

	Duration	Intensity	Pitch
piecewiseSEM	0.546	0.450	0.467
r2.corr.mer	0.567	0.476	0.448

The comparison between the two methods yielded only marginal differences. To begin with the pitch values, for which both R^2 calculation methods successfully processed the model, the piecewiseSEM package yielded a value that is 1.9% higher than the value obtained from r2.corr.mer. In contrast, r2.corr.mer yielded slightly higher R^2 values for duration and intensity models, which were 2.1% and 2.6% higher respectively. Similar discrepancies were also observed when the two methods were used to calculate and compare the R^2 values for the second set of acoustic analyses, which are provided in Table 6.6.

Table 6.6

Comparison of R^2 values obtained for Acoustic Analysis 2

	Duration	Intensity	Pitch
piecewiseSEM	0.539	0.441	0.465
r2.corr.mer	0.572	0.483	0.468

The observed differences confirm the earlier discussion on the lack of consensus on which R^2 measure to report or whether to report one at all. That said, because the piecewiseSEM package is specifically designed for mixed-effects models and its syntax for duration and intensity models included the random effects with their corresponding intercepts, which was missing in the case of r2.corr.mer, the goodness-of-fit measures obtained from the piecewiseSEM package are reported in the Results section.

6.7.6. Pairwise comparisons

When the results of the linear mixed-effects analysis are displayed in R, by default the intercept is set to the predicted mean of one of the fixed effects. Once the intercept has been modelled, the predicted means of the remaining effects are compared to this reference level (see Appendix F). In the current study, the predicted mean of the pretest productions by the intervention group was set as the intercept and the predicted means of the three remaining sets of productions, namely, control pretest, intervention posttest and control posttest were compared to the intercept. This would be fine for standard regression reporting, where the changes in the baseline coefficient are observed in relation to other coefficients. However, the design of the current study made it necessary to obtain pairwise comparisons, much like that of a paired-samples t-test style output, in order to be able to analyze the differences between the pretest and posttest productions by the two experimental groups. Releveling the variables *condition* and *time* to change the intercept and producing four different outputs with four different comparisons would have been malpractice as it would have ignored the power adjustment that is necessary for running multiple comparisons. A solution to this issue is using the least-squares means (lsmeans) package version 2.27-2 (Lenth, 2016). The lsmeans package generates a grid based on the fixed effects entered into the lmer model and uses this grid to display the adjusted means of the coefficients in table form (Lenth, 2018). The lsmeans package can also run tukey-adjusted pairwise comparisons based on the fitted lmer model, which provides a solution for the multiple comparison issue inherent to lmer results. The lsmeans formula below was used to obtain the model estimates by experimental condition and time as well as to run the post hoc analysis for the pairwise differences of contrast. The same lsmeans formula was used for the analysis of duration, intensity, and pitch measures in the case of both sets of acoustic analyses. It should be noted that the component *lmermodel* in the formula below is representational. In the actual analysis, an object that has been assigned a particular mixed-effects model (e.g., the model fitted for the analysis of duration measures) substituted for *lmermodel*.

lsmeans (lmermodel, list (pairwise ~ time|condition), adjust= “tukey”)

For the second set of acoustic analyses, the lsmeans formula above was modified by substituting the fixed effect *feedback* for *condition.*

lsmeans(lmermodel, list(pairwise ~ time|feedback), adjust= “tukey”)

Finally, the significance level was set at .05 for all analyses.

6.8. Analysis of listener judgments

Listeners’ judgments of stress placement that had been automatically saved into .csv files were loaded onto R for the analysis of interrater reliability. A number of issues were considered when choosing the test for the analysis of interrater reliability such as the number of raters, the type of data, and the purpose of analysis (Gisev, Bell, & Chen, 2013; Stemler, 2004). In the current study, there were 10 raters and the scale used for the judgment of stress placement was nominal with three levels of measurement: Syllable 1, Syllable 2, and Syllable 3. Therefore, a Fleiss’ kappa analysis was run to analyze the interrater reliability of the judgments on stress placement. Also, a series of unweighted Cohen’s kappa was run for all possible rater pairings as a post hoc analysis.

CHAPTER 7: RESULTS

7.1. Acoustic analysis 1

The first set of acoustic analyses focused on the pretest and the posttest realizations of syllable duration, intensity, and pitch by the control and the intervention groups. The changes in each one of the three acoustic correlates were analyzed separately.

7.1.1. Duration

The first analysis tested the degree to which recasts had an impact on second syllable duration by comparing the pretest and the posttest measures for proportional difference in duration. According to the descriptive statistics, the intervention group produced second syllables with longer duration on the posttest than they did on the pretest with a difference of 3.6%, whereas there was no difference (0%) between the pretest and posttest productions of the control group (see Table 7.1).

Table 7.1

Descriptive statistics for proportional difference in duration by experimental condition

Group	Time	*M*	*SD*	*Minimum*	*Maximum*
Control	Pretest	−0.242	0.153	−0.719	0.270
	Posttest	−0.242	0.151	−0.666	0.168
Intervention	Pretest	−0.269	0.145	−0.653	0.168
	Posttest	−0.233	0.150	−0.565	0.222

Next, the mixed-effects analysis was run for parametric statistics. The model fitted for duration yielded a significant main effect for condition with the control group producing the second syllables of the target words with longer duration on the pretest than the intervention group. There was also a significant main effect of time with the positive estimate indicating that the intervention group produced the second syllables of the target words with longer duration on the posttest. The effect of time was larger than the effect of group. The model explained 54.6% of the variance, which is considered to be a medium effect size. As mentioned in Chapter 6.7.6, the output for linear mixed-effects analysis does not produce pairwise comparisons. Although Table 7.2 shows that the intervention group produced the second syllables with relatively longer duration on the posttest, it does not show how the posttest performance of the control group compares to their pretest performance. Table 7.3, on the other hand, provides further details on the model estimates obtained through the lsmeans formula used for the post hoc analysis. The results from the post hoc analysis indicated that in the case of the control group, there was no difference between the pretest and the posttest measures of proportional difference in duration. On the other hand, the intervention group produced second syllables with 3.6% longer duration on the posttest compared to their productions on the pretest. Further details on pairwise differences for both experimental groups are shown in Table 7.4 and Figure 7.1 which provides a graphical representation of the confidence intervals. According to pairwise differences of contrasts, there was no statistical significance between the pretest and

Table 7.2

Linear mixed-effects model estimates for proportional difference in duration by experimental condition

Formula: lmer (duration ~ condition * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

	β estimate	*SE*	*df*	*t*-value	*p*-value
Intercept	−0.276	0.025	22.80	−11.244	.000
ConditionControl	0.026	0.012	67.60	2.046	.045
TimePosttest	0.036	0.008	443.60	4.336	.000
ConditionControl: TimePosttest	−0.037	0.012	804.40	−3.022	.003

Goodness of fit: Conditional R^2 = 0.546

Table 7.3

Predicted pretest and posttest means for proportional difference in duration by experimental condition

Group	Time	β estimate	*SE*	*df*	lower *CL*	upper *CL*
Control	Pretest	−0.251	0.025	23.66	−0.302	−0.199
	Posttest	−0.251	0.026	22.62	−0.304	−0.198
Intervention	Pretest	−0.276	0.025	22.85	−0.327	−0.225
	Posttest	−0.240	0.025	21.94	−0.293	−0.187

Table 7.4

Pairwise pretest-posttest contrasts for proportional difference in duration by experimental condition

Group	β estimate	*SE*	*df*	*t*-value	*p*-value
Control	−0.001	0.009	506.34	−0.063	.950
Intervention	0.036	0.008	443.59	4.336	.000

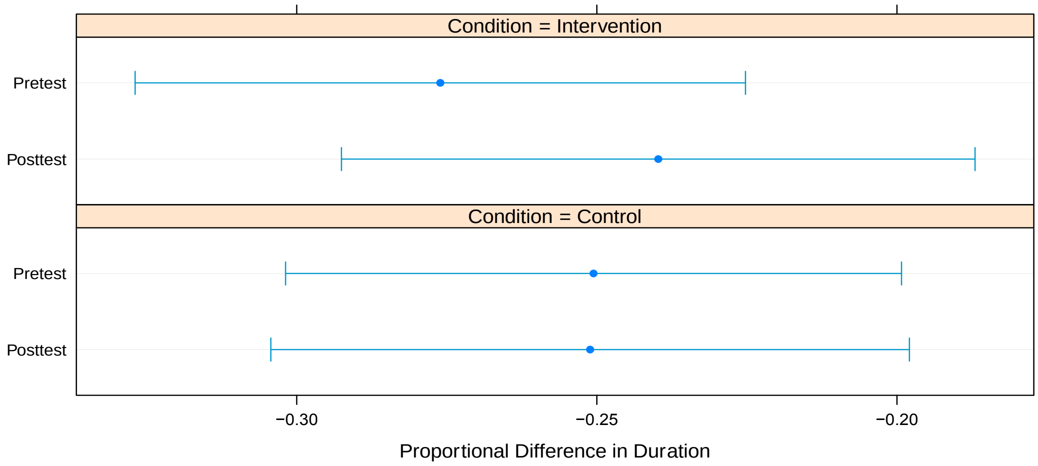


Figure 7.1 95% confidence intervals for the proportional difference in duration by experimental condition

the posttest productions of the control group. However, the gains by the intervention group were statistically significant.

In addition to the analysis of the overall patterns, learner-level and word-level patterns were also inspected. To do this, boxplots were created with the ggplot2 package (Wickham, 2016) using the predicted values from the fitted linear-mixed effects models. The purpose of analyzing boxplots was to understand whether the collective patterns are replicated at the individual learner and word levels consistently. The visual representation of learner-level data yielded a clear pattern simulating the patterns observed in group-level data for both experimental conditions. To begin with the intervention group, boxplots in Figure 7.2 show that every learner in the intervention group produced the second syllables of the target words with relatively longer duration on the posttest in response to the recasts they received. Although there is some variation among learners in the intervention group in terms the amount of gains they have achieved, the overall pattern for gains is clear. On the other hand, the pattern that emerged among the control group learners showed that there was no effect of time and their production of syllable duration did not change on the posttest. In fact, the boxplots in Figure 7.3 show that control group learners' pretest and posttest productions were almost identical. Patterns observed at the learner-level provided further support for the group-level analysis as the results of the model were replicable across almost every learner. Next, a second set of boxplots were created based on the predicted values to inspect the word-level distribution of duration measures. The second set of boxplots also revealed a consistent pattern indicating an increase in proportional duration on the posttest for the words that were produced by the intervention group (Figure 7.4).

The visual representation of the data showed that every target word was

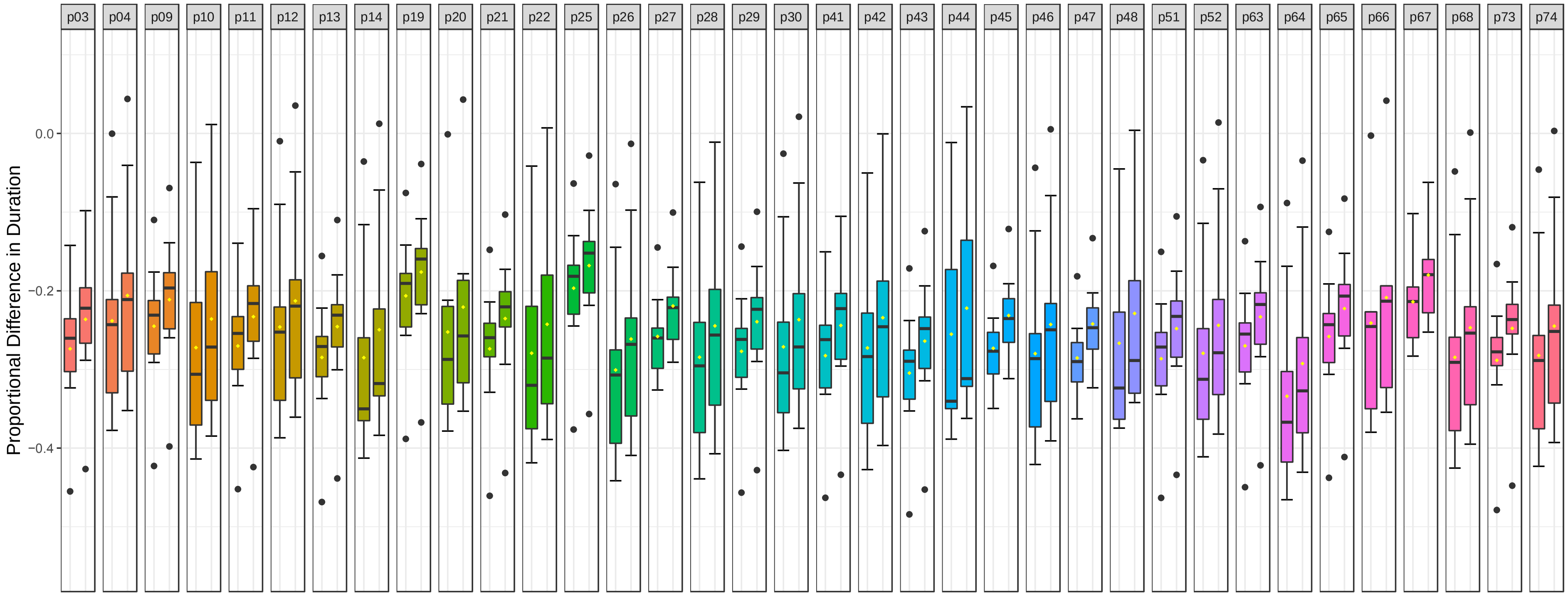


Figure 7.2 Pretest (left) and posttest (right) distributions of duration measures by learners in the intervention group

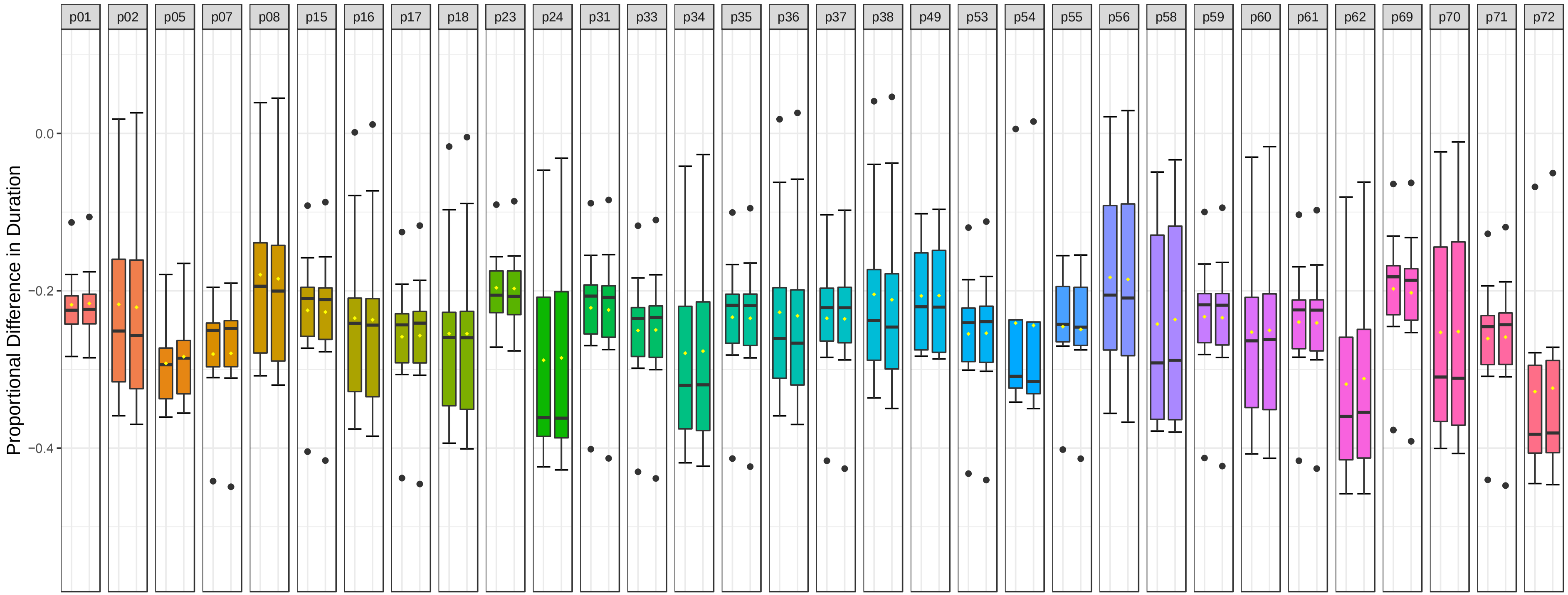


Figure 7.3 Pretest (left) and posttest (right) distributions of duration measures by learners in the control group

realized with longer second syllable duration on the posttest. There was some variation among different target words in terms of the amount of gains they exhibited on the posttest. For example, the words *companion, compassion, consensus, consider,* and *convention* were produced with greater second syllable duration on the posttest in comparison to words such as *condition, diminish, duration,* or *suspicion*. That said, the level of variation did not undermine the robust pattern that provides further support for the consistency of the data as well as the effects of recasts. When it comes to individual words produced by the control group learners, their pretest and posttest productions had highly similar distributions across all target words (Figure 7.5). Although some target words were produced with longer second syllable duration on the posttest (e.g., compassion, consider, position), there were also words that were produced with shorter second syllable duration (e.g., consensus, contention, duration) as well as words with quite similar pretest and posttest realizations.

The distribution patterns observed at the word-level is highly similar to learner-level and group-level data in that they all indicate gains on the posttest by the intervention group, whereas the pretest and posttest distributions of duration measures by the control group were highly stable. In other words, the word-level data, the learner-level data, and the group-level data are all aligned for both experimental conditions, which provides further support for the effects of recasts on increased second syllable duration.

One last set of boxplots were created to compare learners' production of syllable duration to the native-speaker baseline. The learner data used in the boxplots was the predicted values from the linear mixed-effects model and the native-speaker data was based on descriptive statistics. Once again, the pretest and posttest productions by the control group looked highly similar, whereas the intervention

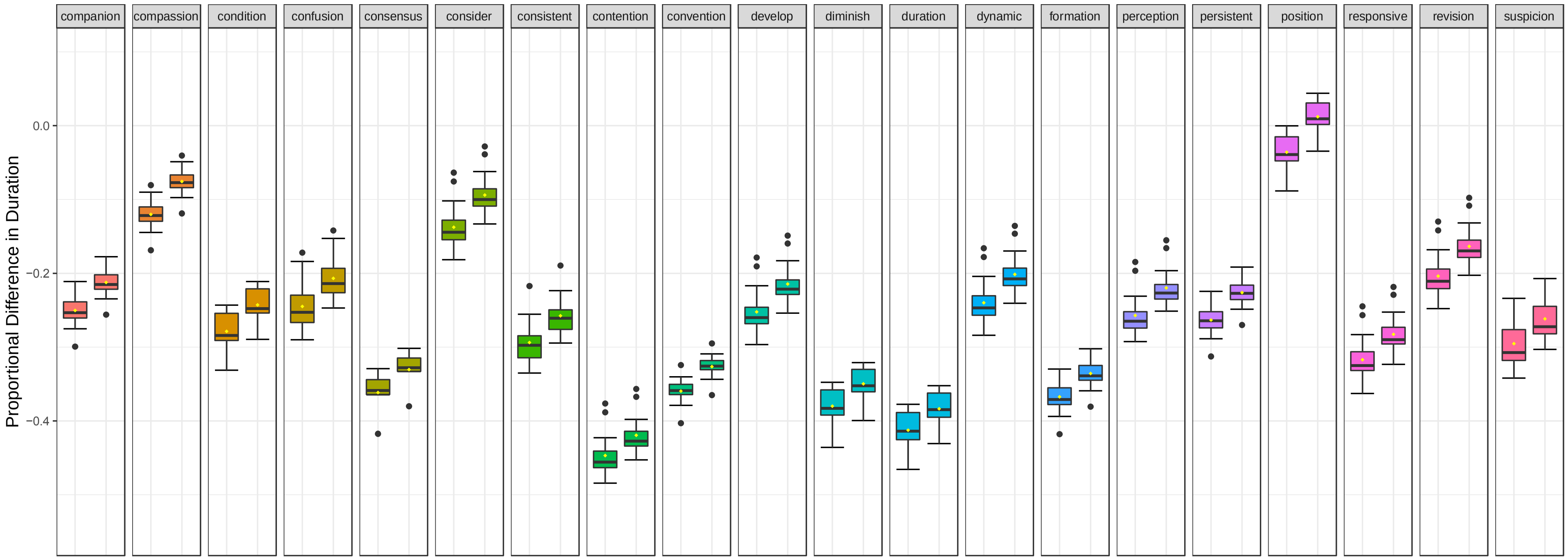


Figure 7.4 Pretest (left) and posttest (right) distributions of duration measures at the word level by learners in the intervention group

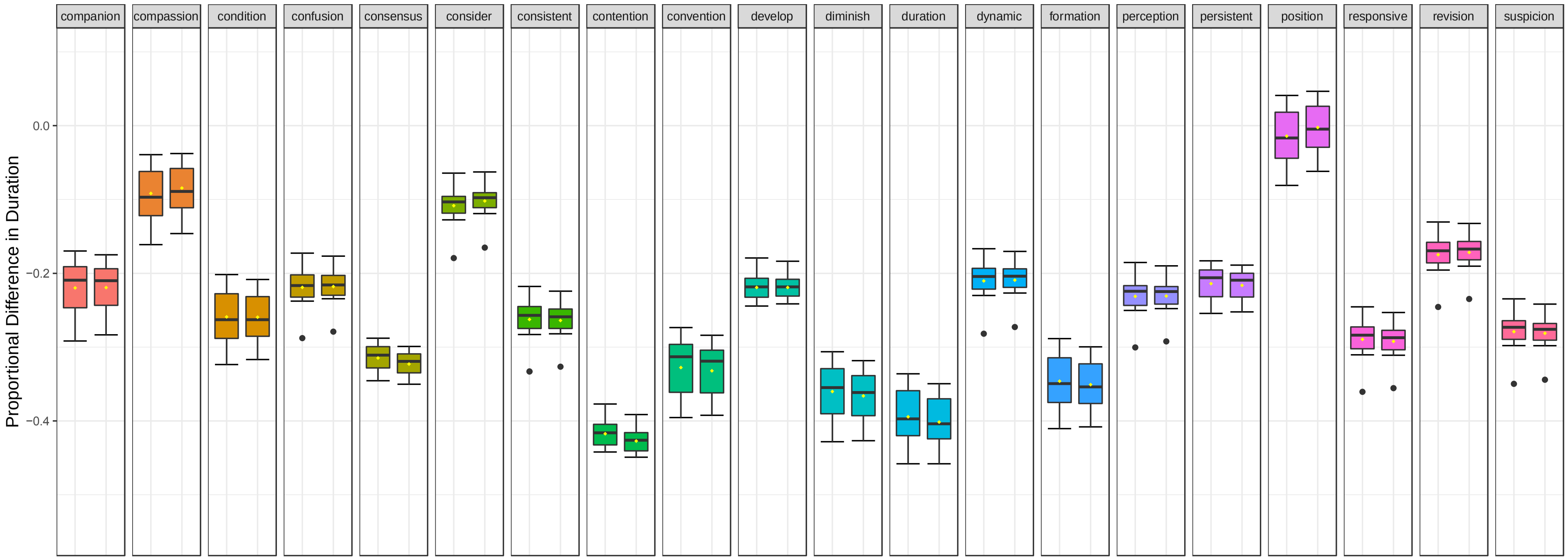


Figure 7.5 Pretest (left) and posttest (right) distributions of duration measures at the word level by learners in the control group

group not only produced second syllables with relatively longer duration on the posttest but they also moved in the direction of native-speaker baseline. Comparison to native-speaker baseline provided further insights into the direction of change triggered by recasts and showed that the posttest productions of duration by the intervention group became more target-like. The comparison of learner and native-speaker distributions can be seen in Figure 7.6. The yellow dots indicate the location of the mean for the corresponding boxplots.

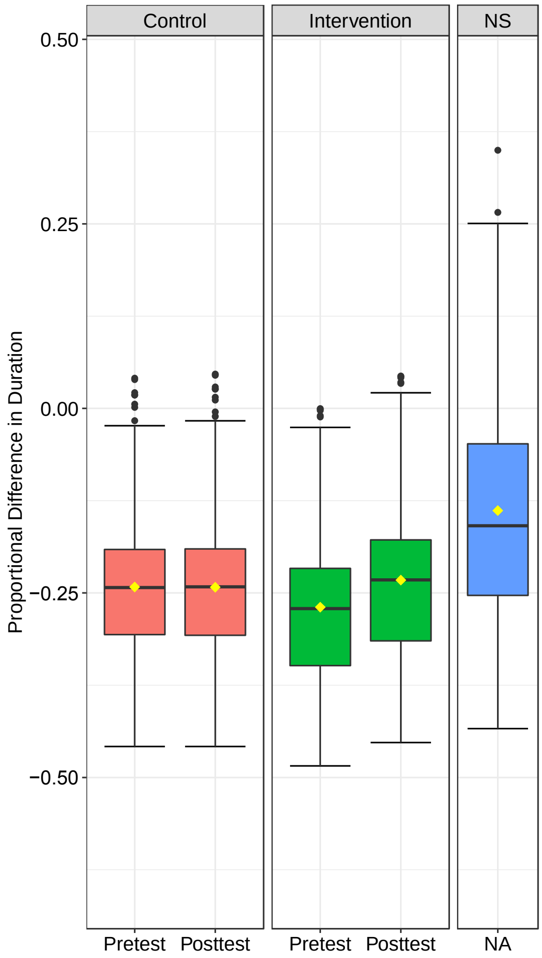


Figure 7.6 Comparison between NS and NNS productions of proportional difference in duration by experimental condition

The intervention group's predicted pretest mean is −0.276 (*SE* = 0.025) whereas their predicted posttest mean is −0.240 (*SE* = 0.025). On the other hand, the mean for the native-speaker baseline is −0.140 (*SD* = 0.160). The differences based on the mean values show that after an increase of 3.6% in proportional difference in duration, there was still a gap of 10% between the intervention group and native-speaker group mean values. In short, the duration of the second syllables produced by the intervention group got closer to native-speaker productions but there was still a gap between the mean values and the overall distributions.

7.1.2. Intensity

The second analysis focused on the changes in proportional difference in intensity. The descriptive statistics provided in Table 7.5 show that there was no difference between the pretest and posttest productions of the control group (0%). In a similar way, the difference was minimal for the intervention group (0.3%).

As the next step, the results of the mixed-effects analysis were inspected. The model did not yield a main effect for condition showing that the pretest productions of both groups were similar. There was no main effect for time either, which means that the intervention group produced the second syllables with similar intensity levels on the pretest and the posttest. The model explained 45% of the variance. This is considered as a medium effect size. Table 7.6 provides an overview of the model estimates for the intensity measure.

In order to analyze the estimated pretest and posttest means of the two experimental conditions, a post hoc analysis was conducted using the lsmeans package. The first part of the post hoc analysis, which is provided in Table 7.7, showed that the control group's pretest and posttest intensity productions were almost the same. On the other hand, the second syllables produced by the intervention group had slightly higher intensity on the posttest compared to the pretest. The second part of the post hoc analysis provided further details on the pairwise comparisons and their

Table 7.5

Descriptive statistics for proportional difference in intensity by experimental condition

Group	Time	*M*	*SD*	*Minimum*	*Maximum*
Control	Pretest	−0.339	0.033	−0.428	−0.255
	Posttest	−0.339	0.035	−0.439	−0.237
Intervention	Pretest	−0.343	0.039	−0.492	−0.229
	Posttest	−0.340	0.035	−0.434	−0.240

Table 7.6

Linear mixed-effects model estimates for proportional difference in intensity by experimental condition

Formula: lmer (intensity ~ condition * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

	β estimate	*SE*	*df*	*t*-value	*p*-value
Intercept	−0.343	0.006	22.30	−61.137	0.000
ConditionControl	0.004	0.003	79.60	1.500	0.137
TimePosttest	0.002	0.002	384.10	1.052	0.294
ConditionControl: TimePosttest	−0.002	0.003	878.50	−0.735	0.463

Goodness of fit: Conditional R^2 = 0.450

Table 7.7

Predicted pretest and posttest means for proportional difference in intensity by experimental condition

Group	Time	β estimate	*SE*	*df*	lower *CL*	upper *CL*
Control	Pretest	−0.339	0.006	23.10	−0.350	−0.327
	Posttest	−0.339	0.005	24.26	−0.350	−0.328
Intervention	Pretest	−0.343	0.006	22.27	−0.354	−0.331
	Posttest	−0.341	0.005	23.23	−0.351	−0.330

statistical significance. According to pairwise differences of contrast, there was no statistically significant difference between the pretest and the posttest productions of the control group. Similarly, the pretest and the posttest productions of intensity by the intervention group did not reach statistical significance. In summary, receiving recasts on target vocabulary did not lead to realization of those words with higher intensity. Table 7.8 shows pairwise comparisons for proportional difference in intensity and Figure 7.7 provides a graphical view of the confidence intervals.

Table 7.8

Pairwise pretest-posttest contrasts for the proportional difference in intensity by experimental condition

Group	β estimate	*SE*	*df*	*t*-value	*p*-value
Control	0.000	0.002	450.87	−0.013	.989
Intervention	0.002	0.002	384.06	1.052	.294

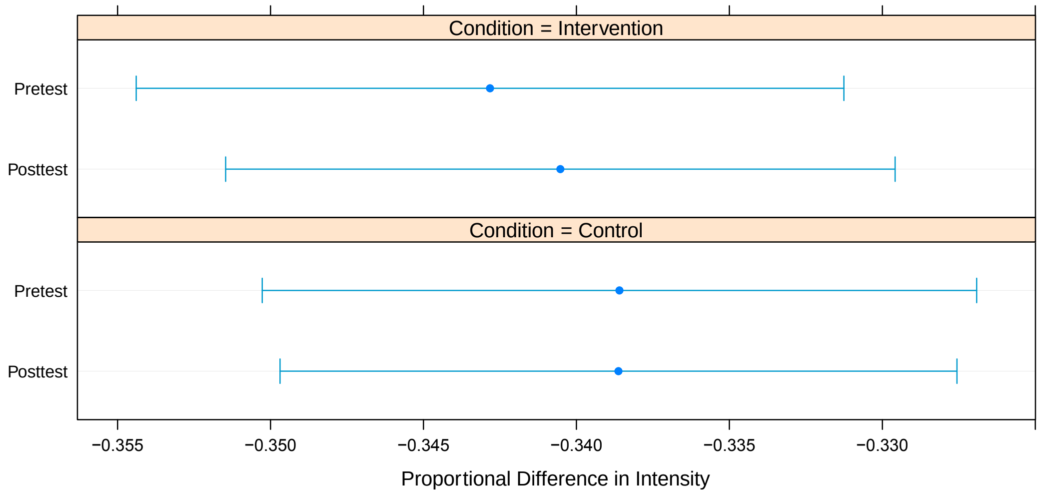


Figure 7.7 95% confidence intervals for proportional intensity difference by experimental condition

Next, in order to inspect the learner-level and the word-level distributions of intensity measures, boxplots were created based on the predicted values. Figures 7.8 and 7.9 provide a visual representation of learner-level distributions for the intensity measures. To begin with the intervention group, the boxplots indicated that a few learners produced second syllables with slightly higher intensity on the posttest. However, despite some variation across individual learners, the differences between the pretest and the posttest measures were minimal for the most part. Likewise, the learners in the control group did not exhibit any gains on the posttest. Taken together, the learner-level distributions did not raise any red flags and they were aligned with

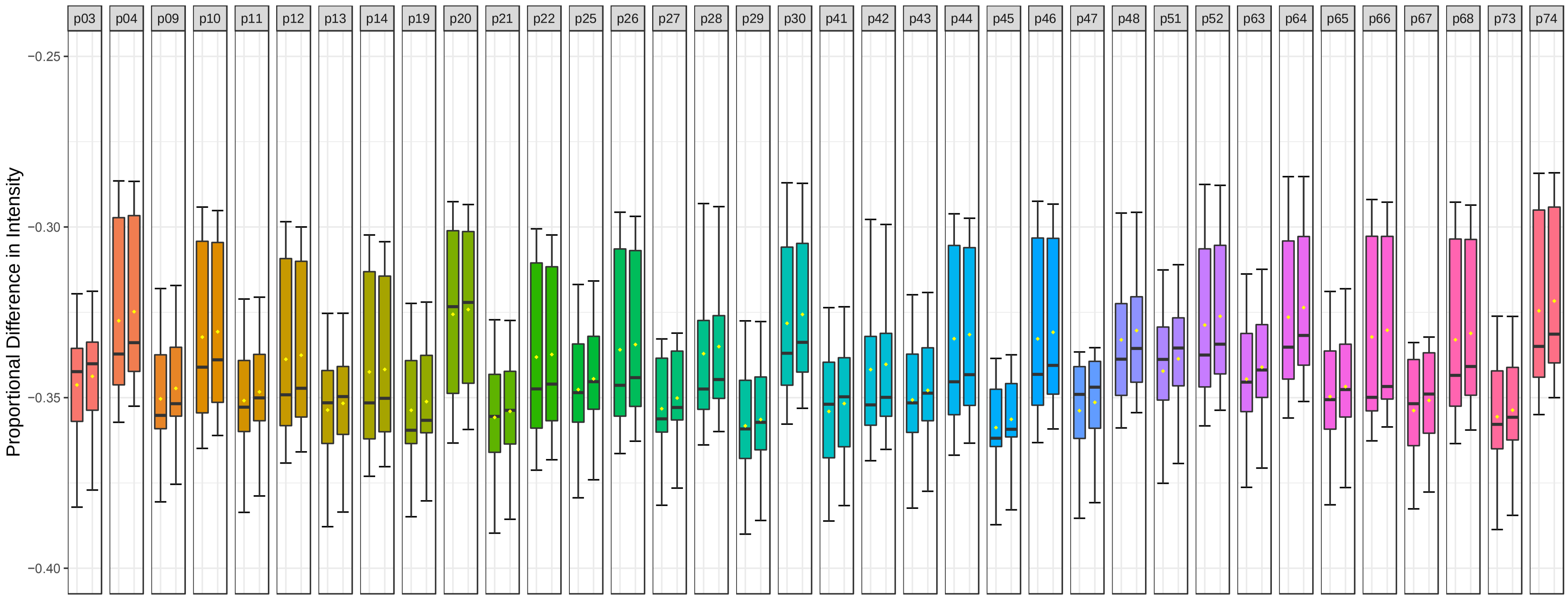


Figure 7.8 Pretest (left) and posttest (right) distributions of intensity measures by learners in the intervention group

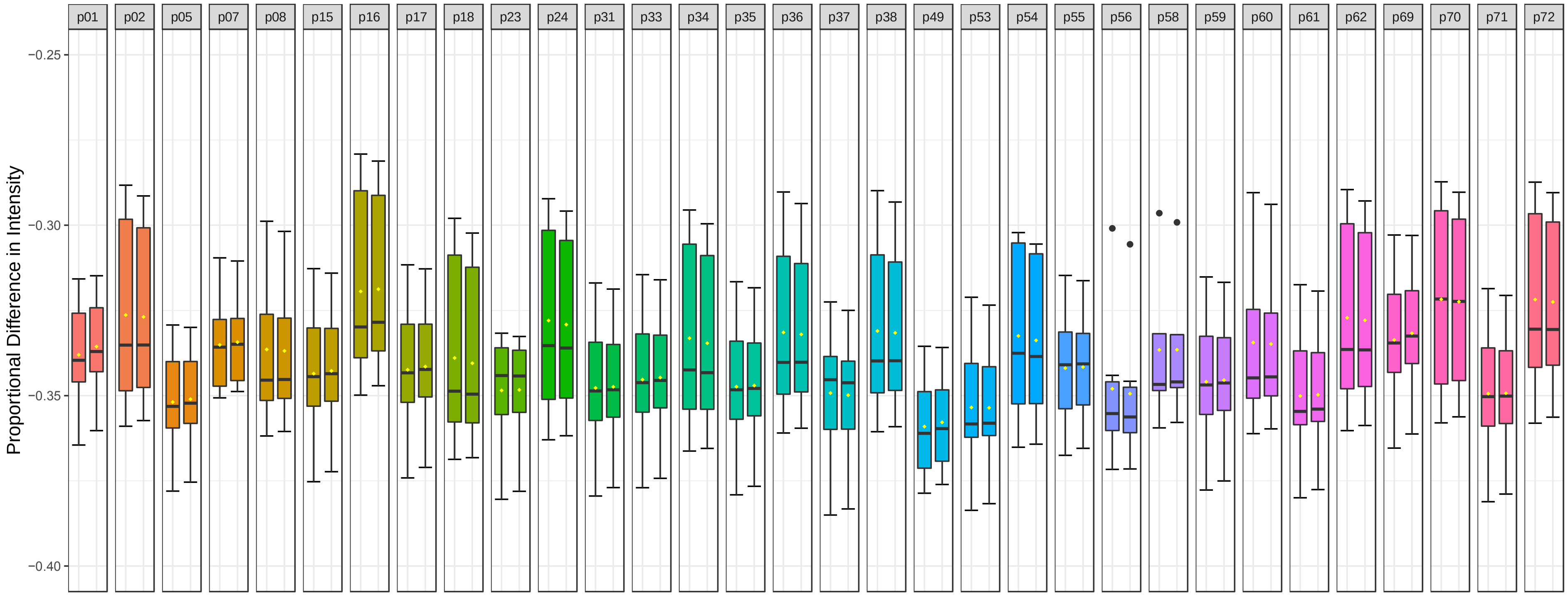


Figure 7.9 Pretest (left) and posttest (right) distributions of intensity measures by learners in the control group

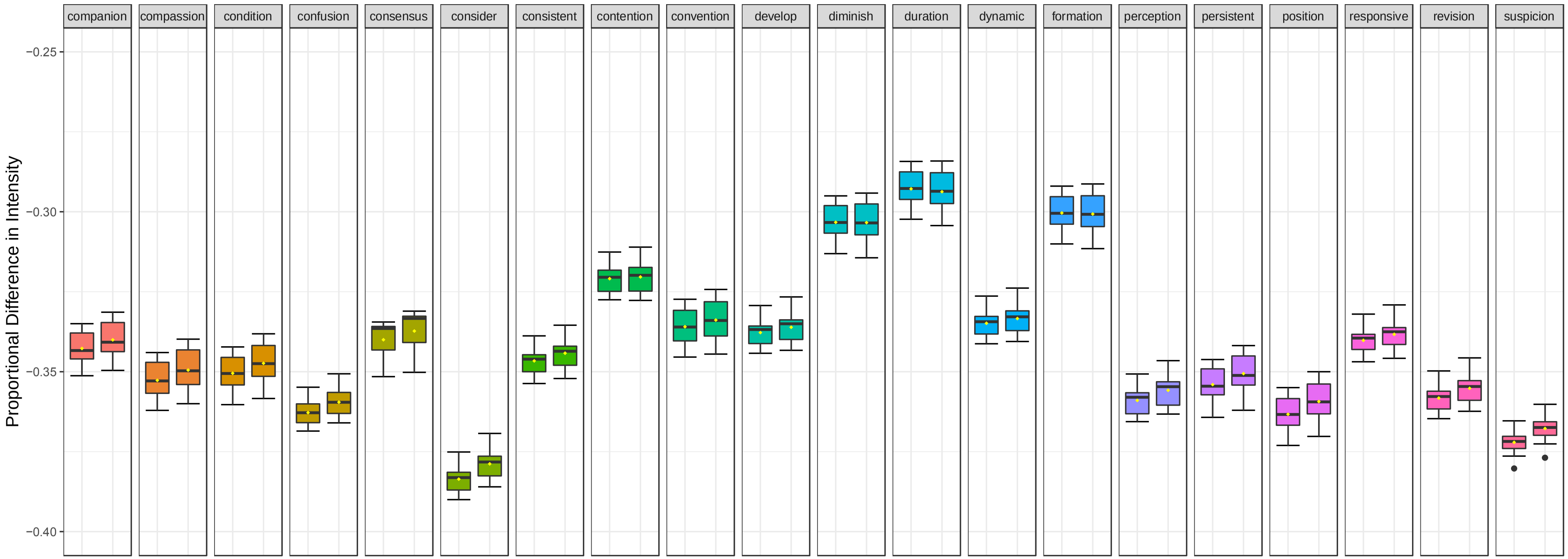


Figure 7.10 Pretest (left) and posttest (right) distributions of intensity measures at the word level by learners in the intervention group

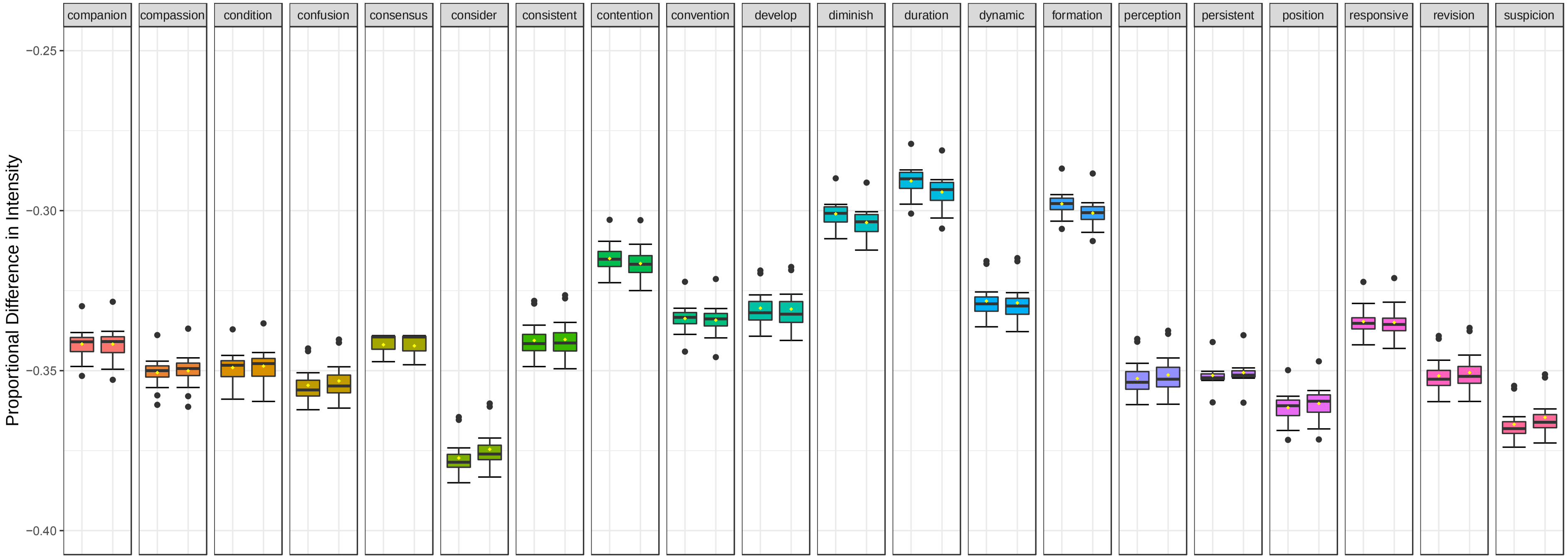


Figure 7.11 Pretest (left) and posttest (right) distributions of intensity measures at the word level by learners in the control group

the results of the overall mixed-effects model for intensity.

When it comes to the word-level distributions of intensity measures, once again the intervention group produced target words with similar second syllable intensity levels on the pretest and the posttest as shown in Figure 7.10. Although some words were produced with higher intensity on the posttest, this was not the general pattern. For the majority of the words, the pretest-posttest differences in proportional intensity were highly marginal compared to much more noticeable differences that were observed in the case of duration measures displayed earlier in Figure 7.4. Also, as can be seen in Figure 7.11, the intensity levels of second syllables produced by the control group did not exhibit any changes on the posttest either. In fact, some words were realized with lower syllable 2 intensity on the posttest. However, once again the differences were minimal. The group-level, learner-level, and word-level distributions of intensity measures were all aligned with no inconsistent patterns in the data.

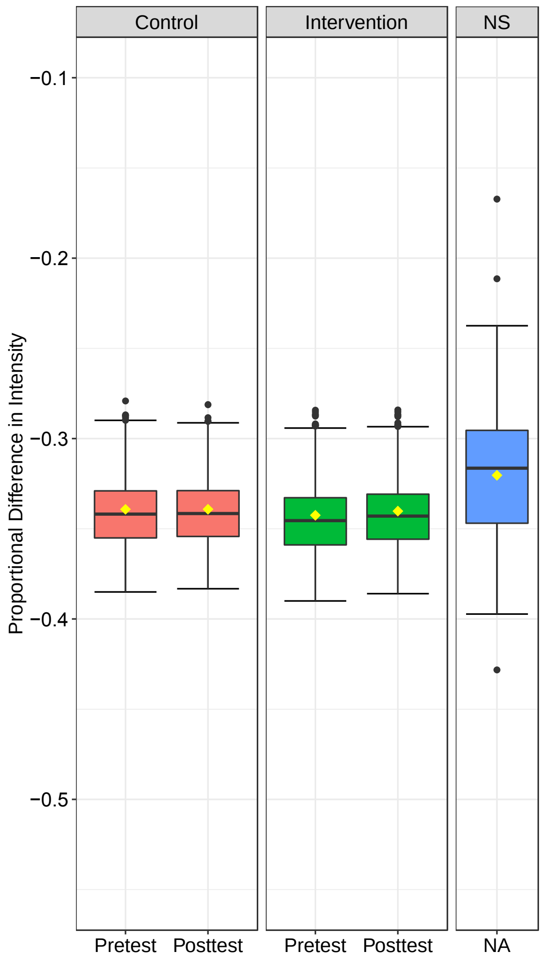


Figure 7.12 Comparison between NS and NNS productions of proportional difference in intensity by experimental condition

As the last step, the pretest and the posttest realizations of syllable intensity by both experimental groups were compared to realizations of syllable intensity by native speakers. According to the comparison between learner and native-speaker productions as shown in

Figure 7.12, both groups of learners produced the second syllables of the target words with less intensity compared to the native-speaker baseline. Also, the learner data was concentrated around the mean whereas the native-speaker productions were more spread out.

7.1.3. Pitch

The third analysis focused on the measures of proportional difference in pitch to examine the extent of the impact that recasts had on the second syllable pitch levels. The descriptive statistics for pitch measures provided in Table 7.9 show that the gains were under 1 percent for both the intervention (0.6%) and the control (0.2%) groups.

Table 7.9

Descriptive statistics for proportional difference in pitch by experimental condition

Group	Time	*M*	SD	*Minimum*	*Maximum*
Control	Pretest	−0.323	0.031	−0.452	−0.193
	Posttest	−0.321	0.031	−0.411	−0.143
Intervention	Pretest	−0.325	0.028	−0.402	−0.203
	Posttest	−0.319	0.031	−0.377	−0.193

The estimates by the linear mixed-effects model fitted for the analysis of pitch measures were explored next. The model did not produce a main effect for condition as the pretest productions of syllable pitch by the two groups were similar. That said, there was a main effect of time, indicating that the second syllables produced by the intervention group had higher pitch values on the posttest. Although statistically significant, it should be noted that the difference was less than 1%. The model explained 46.7% of the variance, which is considered as a medium effect size. An overview of the model estimates for the pitch measures can be seen in Table 7.10. In

Table 7.10

Linear mixed-effects model estimates for proportional difference in pitch by experimental condition

Formula: lmer (pitch ~ condition * time + (1+time | learner) + (1 + time | word))

	β estimate	*SE*	*df*	*t*-value	*p*-value
Intercept	−0.326	0.004	45.22	−90.774	0.000
ConditionControl	0.001	0.004	68.38	0.377	0.708
TimePosttest	0.007	0.002	164.14	3.277	0.001
ConditionControl: TimePosttest	−0.005	0.003	183.84	−1.557	0.121

Goodness of fit: Conditional $R^2 = 0.467$

Table 7.11

Predicted pretest and posttest means for proportional difference in pitch by experimental condition

Group	Time	β estimate	*SE*	*df*	lower *CL*	upper *CL*
Control	Pretest	−0.324	0.004	49.76	−0.332	−0.317
	Posttest	−0.322	0.004	68.37	−0.330	−0.314
Intervention	Pretest	−0.326	0.004	45.22	−0.333	−0.318
	Posttest	−0.319	0.004	64.17	−0.327	−0.311

order to obtain the estimated pretest and posttest means of the two groups with their corresponding confidence intervals, post hoc analyses were conducted using the lsmeans package. As can be seen in Table 7.11, the control group’s the posttest and the pretest productions of syllable pitch were almost at the same level. On the other hand, the intervention group produced second syllables with slightly higher pitch on the posttest compared to their productions of pitch on the pretest.

Next, pairwise contrasts were inspected for the test of significance. Pairwise analyses yielded no statistical difference between the pretest and posttest pitch productions of the control group. However, in the case of the intervention group, the increase in pitch levels on the posttest was statistically significant. According to the results, recasts had an impact on the production of syllable pitch; however, the degree of the effect was not as large as the effect of recasts on syllable duration. The pairwise differences in pitch values and confidence intervals are provided in Table 7.12 and Figure 7.13 respectively.

Finally, the analysis of learner-level and word-level distributions of proportional difference in pitch were carried out using boxplots. To begin with the learner-level distributions, it was seen that there was considerable variation among learners in the intervention group in terms of their gains on the posttest (Figure 7.14).

Table 7.12

Pairwise pretest-posttest contrasts for proportional difference in pitch by experimental condition

Group	β estimate	*SE*	*df*	*t*-value	*p*-value
Control	0.002	0.002	174.92	0.931	.353
Intervention	0.007	0.002	164.14	3.277	.001

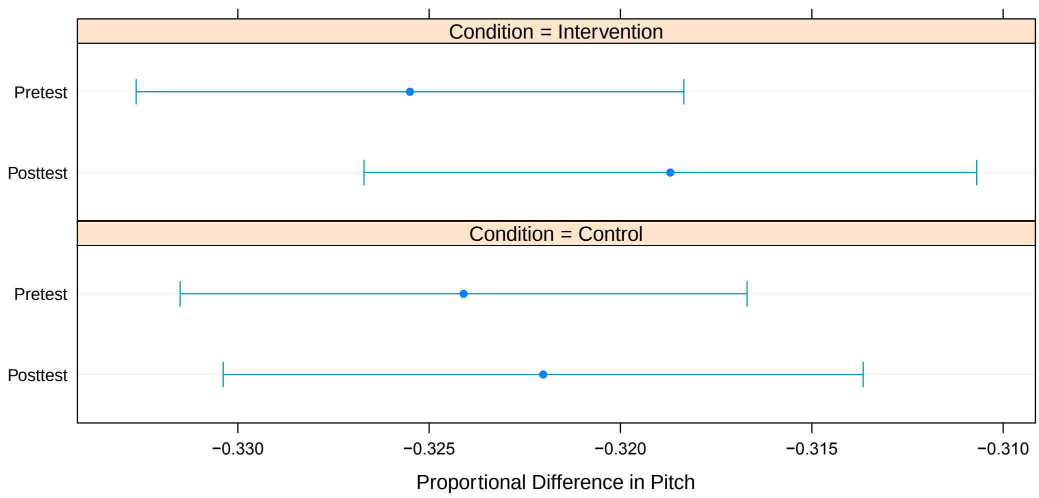


Figure 7.13 95% confidence intervals for proportional difference in pitch by experimental condition

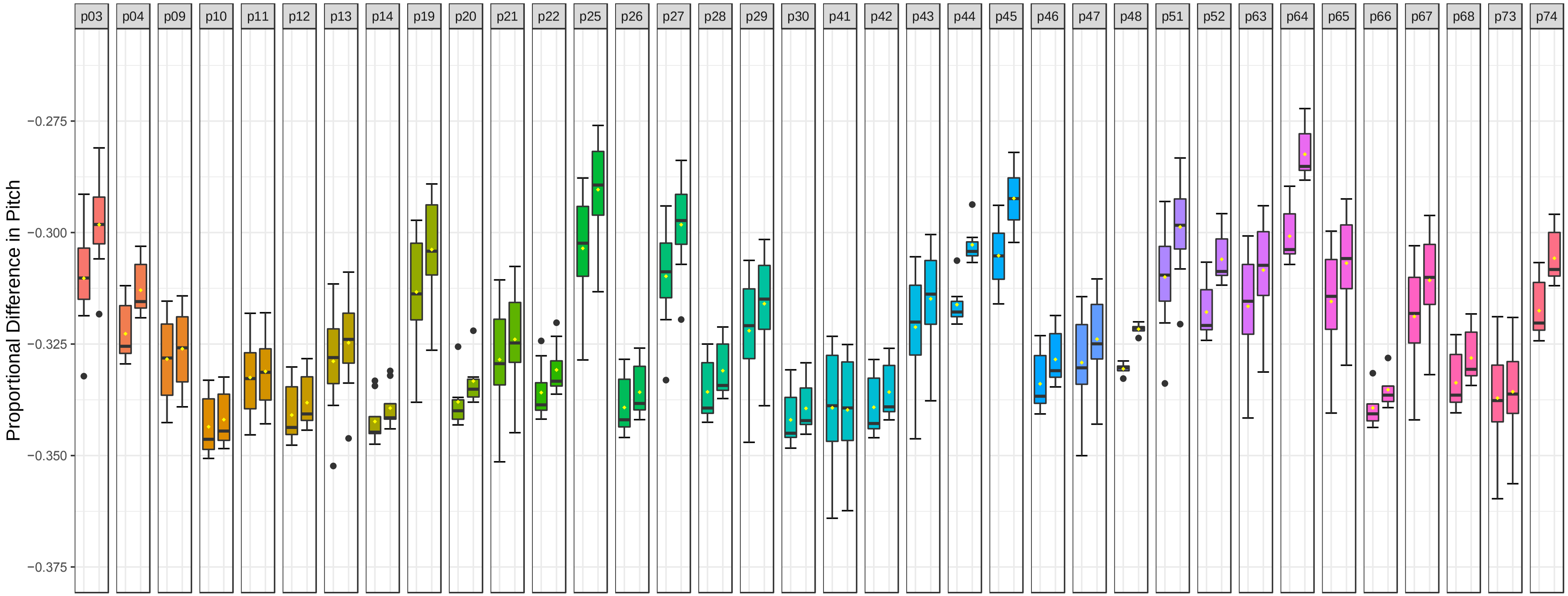


Figure 7.14 Pretest (left) and posttest (right) distributions of pitch measures by learners in the intervention group

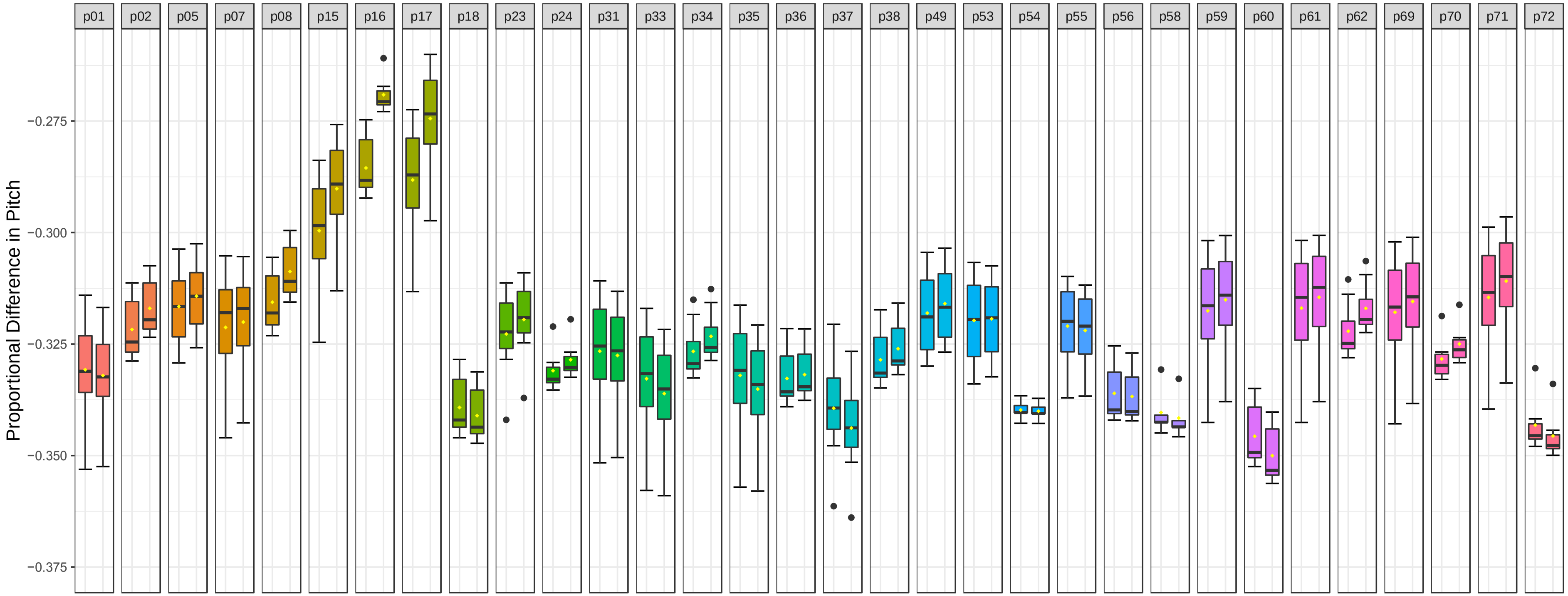


Figure 7.15 Pretest (left) and posttest (right) distributions of pitch measures by learners in the control group

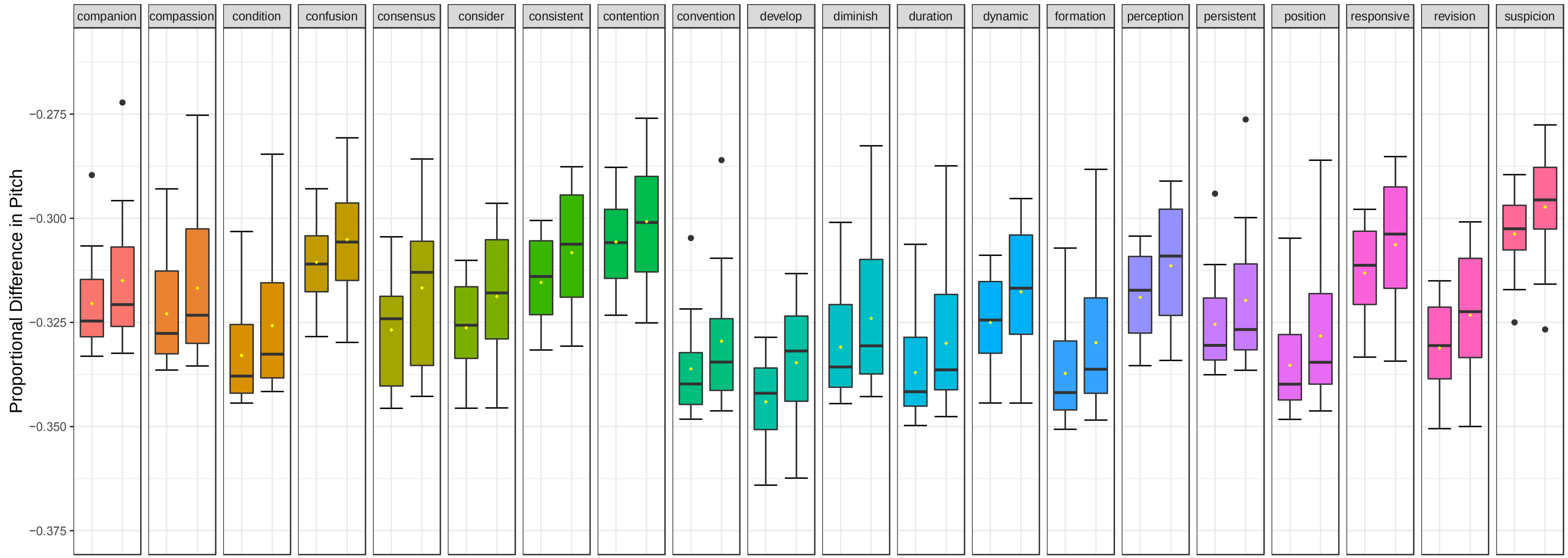


Figure 7.16 Pretest (left) and posttest (right) distributions of pitch measures at the word level by learners in the intervention group

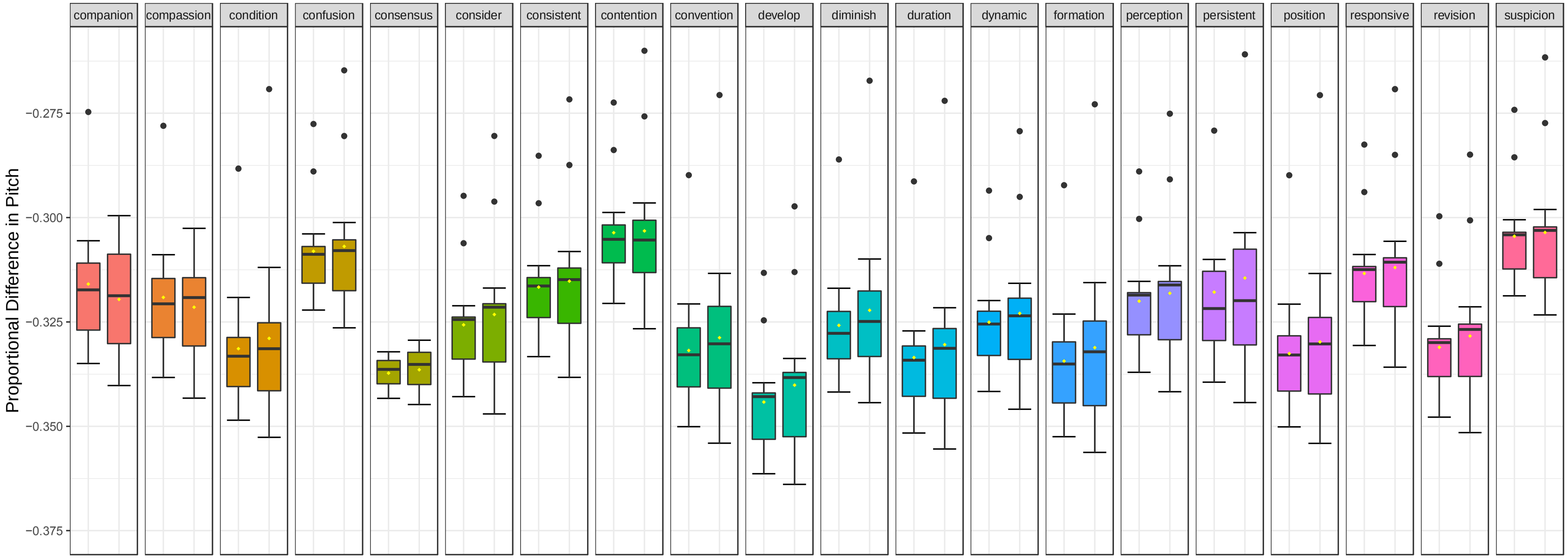


Figure 7.17 Pretest (left) and posttest (right) distributions of pitch measures at the word level by learners in the control group

That is to say, some learners achieved major gains on the posttest, whereas others achieved moderate or minimal gains. The pattern observed in the control group included more variation with fewer learners displaying visible gains and more learners displaying minimal gains as well as learners who produced syllables at a relatively lower pitch on the posttest (Figure 7.15). Overall, the majority of learners in the intervention group produced syllables with higher pitch on the posttest. Yet, the amount of variation among the intervention group learners as well as the control group learners was noticeably higher compared to the patterns observed in the case of duration and intensity measures, which were more stable across both groups of learners. The word-level distributions of pitch measures displayed in Figure 7.16 and Figure 7.17 also confirmed that the words produced by the intervention group had a higher syllable pitch on the posttest. On average, the amount of overlap between pretest and posttest distributions of pitch measures by the intervention group was less than the overlap between pretest and posttest distributions of the pitch levels produced by the control group. In fact, a word-by-word comparison between the two experimental conditions shows that the majority of the target words had a higher syllable pitch on the posttest when they were produced

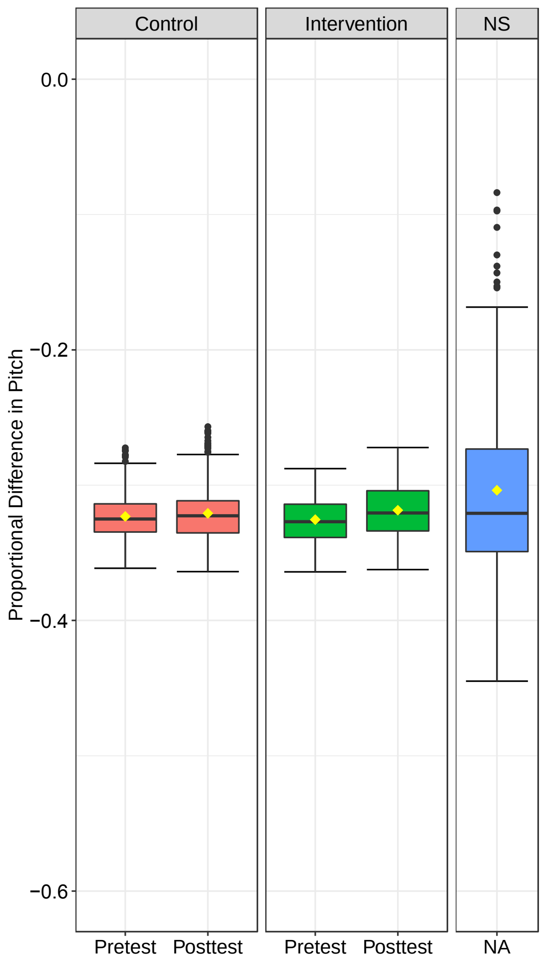


Figure 7.18 Comparison between NS and NNS productions of proportional difference in pitch by experimental condition

by the intervention group.

The final set of boxplots provided a comparison between the native-speaker baseline and the predicted pitch values produced by the two experimental groups. As Figure 7.18 shows, the control group's pretest distributions were almost identical to their posttest distributions, whereas the intervention group showed some gains on the posttest. Although the mean score for posttest productions by the intervention group moved in the direction of native-speaker mean, the native-speaker realizations of pitch have a wider distribution that captures pretest and posttest productions of both experimental conditions. In other words, the gap between learner productions of pitch and the native-speaker baseline was much less than the gap that was observed in the case of duration measures.

7.2. Acoustic analysis 2

7.2.1. Duration

As part of the second set of analyses, the effects of recasts on relative second syllable duration were analyzed by comparing the target words that received a recast and the target words that did not. To begin with descriptive statistics, the mean proportional difference in duration increased by 0.1% on the posttest for the words that did not receive a recast, whereas the difference was 5.7% for the words that received a recast. As it is shown in Table 7.13, the pretest-posttest difference was much larger for the words that received a recast. The mixed-effects model fitted with feedback as a fixed effect provided further details on the gains achieved by the two groups of words. As can be seen in Table 7.14, the model revealed a significant main effect for feedback with the words that received a recast being produced with relatively shorter second syllable duration on the pretest than the words that did not

Table 7.13

Descriptive statistics for proportional difference in duration by feedback provision

Feedback	Time	*M*	SD	*Minimum*	*Maximum*
No Recast	Pretest	−0.246	0.146	−0.527	0.168
	Posttest	−0.247	0.154	−0.565	0.217
Recast	Pretest	−0.282	0.144	−0.653	0.087
	Posttest	−0.225	0.148	−0.527	0.223

Table 7.14

Linear mixed-effects model estimates for proportional difference in duration by feedback provision

Formula: lmer (duration ~ feedback * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

	β estimate	*SE*	*df*	*t*-value	*p*-value
Intercept	−0.289	0.023	22.90	−12.399	0.000
FeedbackNoRecast	0.036	0.013	549.50	2.827	0.005
TimePosttest	0.055	0.011	133.70	5.209	0.000
FeedbackNoRecast: TimePosttest	−0.057	0.017	551.20	−3.305	0.001

Goodness of fit: $R^2 = 0.539$

receive a recast. Furthermore, there was a significant main effect of time with the positive estimate revealing that the intervention group produced the second syllables of the target words with longer duration on the posttest. Once again, the effect of time was larger than the effect of group.

The goodness-of-fit calculation yielded a medium effect size, which showed that the model explained 53.9% of the variance. Post hoc analyses were conducted in order to inspect the pretest and the posttest means of the two groups of words. As shown in Table 7.15, there was less than 1% difference in relative duration measures between the pretest and posttest realizations of the words that did not receive a recast.

Table 7.15

Predicted pretest and posttest means for proportional difference in duration by feedback provision

Feedback	Time	β estimate	*SE*	*df*	lower *CL*	upper *CL*
No Recast	Pretest	−0.252	0.024	27.34	−0.302	−0.202
	Posttest	−0.254	0.027	25.08	−0.310	−0.198
Recast	Pretest	−0.288	0.023	22.86	−0.336	−0.241
	Posttest	−0.233	0.026	21.66	−0.287	−0.179

Table 7.16

Pairwise pretest-posttest contrasts for proportional difference in duration by feedback provision

Feedback	β estimate	SE	df	t-value	p-value
No Recast	−0.002	0.014	287.77	−0.115	.909
Recast	0.055	0.011	133.73	5.209	.000

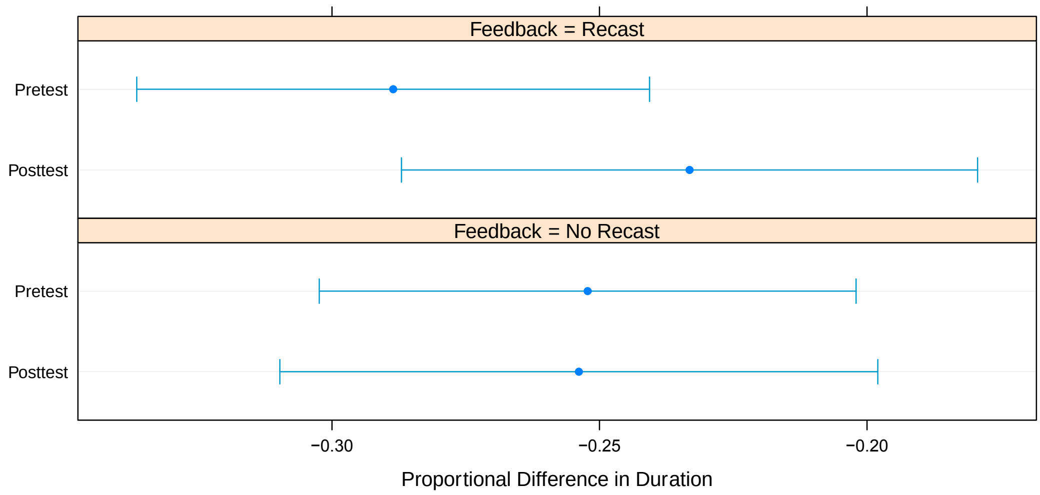


Figure 7.19 95% confidence intervals for proportional difference in duration by feedback provision

On the other hand, the difference was 5.5% between the pretest and the posttest realization of the words that received a recast.

When pairwise differences of contrasts were examined, it was seen that the difference over time was not statistically significant in the case of the words that did not receive a recast. On the other hand, the 5.5% gains in syllable duration observed in the case of words that received a recast was statistically significant. Table 7.16 shows pairwise contrasts and Figure 7.19 shows confidence intervals in graphical form.

Unlike the first set of analyses for which learner-level and word-level distributions were inspected through boxplots, it was not possible to create those boxplots for the second set of acoustic analyses as it meant dividing the data further into smaller groups. As a result of fewer data points, the boxplots would not have provided the same level of information that the group-level data provided and the interpretations based on those boxplots would have not been meaningful. Thus, only boxplots that compared learners' productions to the native-speaker baseline were created and inspected. As shown in Figure 7.20, the words that did not receive a recast had very similar distributions with similar estimated means on the pretest and the posttest. The posttest productions of

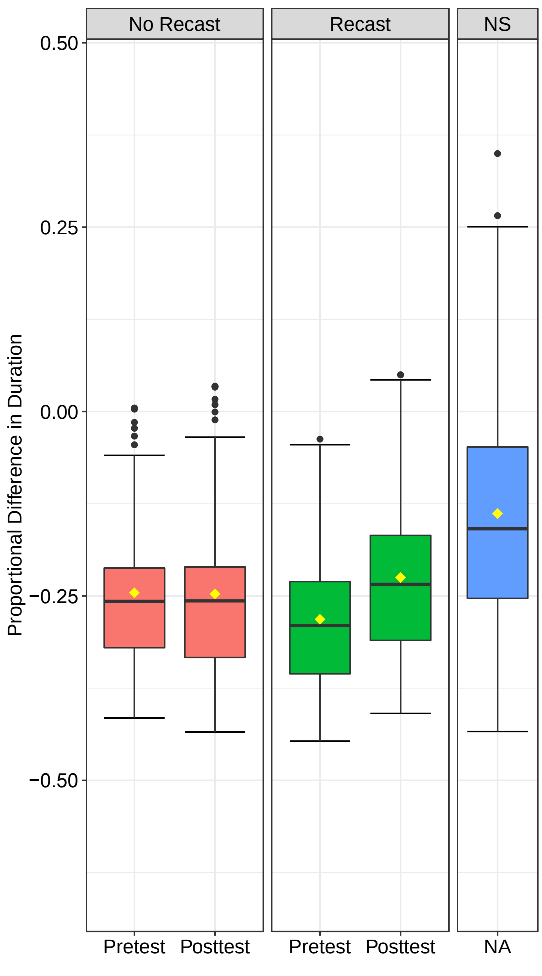


Figure 7.20 Comparison between NS and NNS productions of proportional difference in duration by feedback provision

words that received a recast, on the other hand, had longer syllable duration. In addition, the posttest distribution of the words that received a recast moved in the direction of the native-speaker baseline. These results provided further support for the efficacy of recasts when addressing phonological errors by providing insights into the effects that recasts can potentially have on syllable duration.

7.2.2. Intensity

After the analysis of duration measures by feedback provision, the intensity measures were analyzed. According to the descriptive statistics, there were virtually no differences between pretest and posttest productions of intensity measures for the words that did not receive feedback (0.3%) or the words that did (0.2%). In other words, there were no meaningful differences between the pretest and the posttest realizations of intensity regardless of feedback provision (see Table 7.17).

Table 7.17

Descriptive statistics for proportional difference in intensity by feedback provision

Feedback	Time	*M*	SD	*Minimum*	*Maximum*
No Recast	Pretest	−0.339	0.036	−0.406	−0.229
	Posttest	−0.336	0.032	−0.417	−0.264
Recast	Pretest	−0.344	0.041	−0.492	−0.240
	Posttest	−0.342	0.036	−0.434	−0.240

The inferential statistics based on the mixed-effects model were inspected next (see Table 7.18). The model showed that there was no main effect of feedback when pretest productions of the two groups of words were compared as the second syllables of both groups of words were produced with a similar level intensity. The model did not yield a main effect for time either, which indicated that the recast did not lead to an increase in syllable intensity. Finally, the model had a goodness-of-fit value of

Table 7.18

Linear mixed-effects model estimates for proportional difference in intensity by feedback provision

Formula: lmer (intensity ~ feedback * time + (1| learner) + (0 + time | learner) + (1|word) + (0 + time | word))

	β estimate	*SE*	*df*	*t*-value	*p*-value
Intercept	−0.344	0.006	21.30	−56.057	0.000
FeedbackNoRecast	0.004	0.003	539.80	1.238	0.216
TimePosttest	0.002	0.003	81.00	0.591	0.556
FeedbackNoRecast: TimePosttest	0.001	0.004	549.10	0.303	0.762

Goodness of fit: $R^2 = 0.441$

Table 7.19

Predicted pretest and posttest means for proportional difference in intensity by feedback provision

Feedback	Time	β estimate	*SE*	*df*	lower *CL*	upper *CL*
No Recast	Pretest	−0.340	0.006	25.77	−0.353	−0.327
	Posttest	−0.337	0.006	30.71	−0.348	−0.325
Recast	Pretest	−0.344	0.006	21.30	−0.357	−0.332
	Posttest	−0.342	0.005	23.97	−0.353	−0.332

Table 7.20

Pairwise pretest-posttest contrasts for proportional difference in intensity by feedback provision

Feedback	β estimate	*SE*	*df*	*t*-value	*p*-value
No Recast	0.003	0.004	194.45	0.804	.422
Recast	0.002	0.003	81.01	0.591	.556

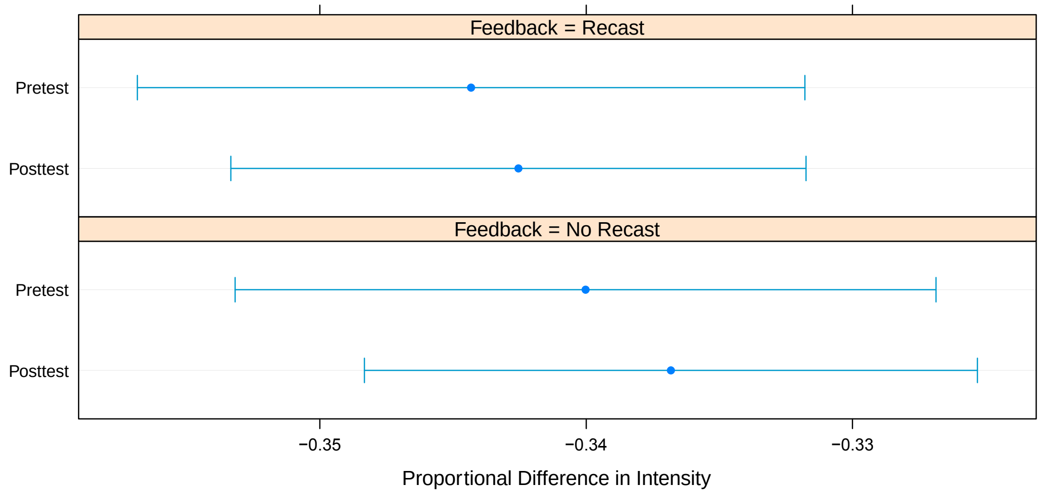


Figure 7.21 95% confidence intervals proportional intensity difference by feedback provision

0.441, which explained less of the variance compared to the model fitted for the analysis of duration.

The post hoc analysis provided in Table 7.19 showed that there was 0.3% difference between the pretest and posttest measures of intensity for the words that did not receive a recast. Similarly, the pretest and posttest difference in intensity measures was 0.2% for the words that received a recast. According to the pairwise analyses, neither the words that did not receive a recast nor the ones that did had statistically higher second syllable intensity on the posttest. These results were aligned with the group-level comparison which showed that recast did not lead to an increase in syllable intensity. The pairwise comparisons can be seen in Table 7.20 and Figure 7.21.

As the last step of the analysis of intensity measures, the distribution of the predicted values for both groups of words were compared to the native-speaker distributions of intensity (Figure 7.22). The boxplots showed that native-speaker baseline for intensity had a higher mean than both pretest and posttest realizations of

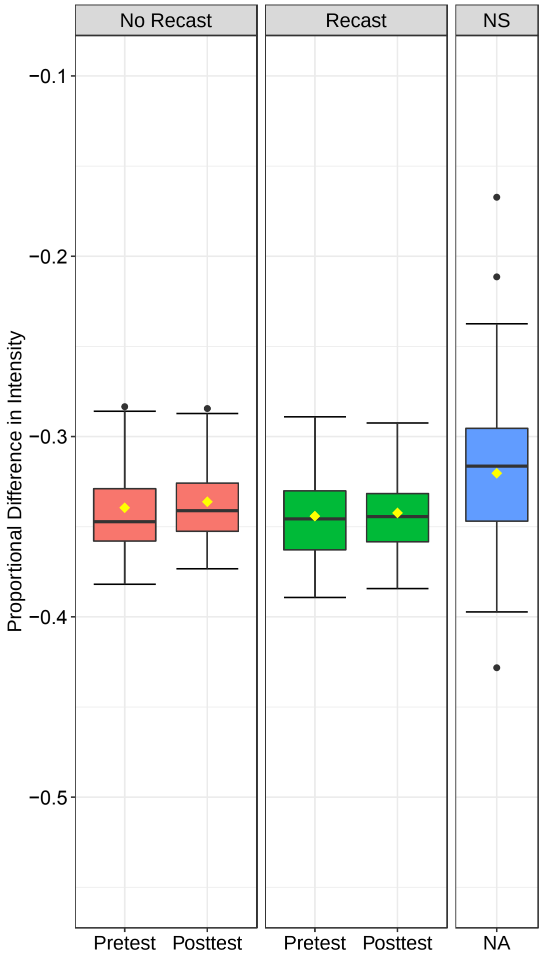


Figure 7.22 Comparison between NS and NNS productions of proportional difference in intensity by feedback provision

the two groups of words. In summary, the analysis of intensity showed that recasts did not have an impact on the production of syllable intensity and realizations of both groups of words had lower intensity than the native-speaker productions.

7.2.3. Pitch

According to descriptive statistics, the pretest and the posttest difference in the mean values for pitch measures was 0.3% for the words that did not receive a recast and 0.9% for the words that received a recast. Table 7.21 provides an overview of the descriptive statistics. As the next step, the output from the mixed-effects model were inspected. The results, which are provided in Table 7.22, showed that there was a main effect of feedback with words that did not receive a recast being produced with higher second syllable pitch on the pretest. The results also showed a main effect for time, which indicated that the words that received a recast were produced with a higher second syllable pitch on the posttest. The model had a goodness-of-fit value of 0.465.

As it has been done previously, post hoc analyses were conducted for obtaining estimated means and pairwise contrasts. The estimated pretest and posttest mean values for both groups of words are provided in Table 7.23.

Table 7.21

Descriptive statistics for proportional difference in pitch by feedback provision

Feedback	Time	*M*	SD	*Minimum*	*Maximum*
No Recast	Pretest	−0.320	0.029	−0.378	−0.237
	Posttest	−0.317	0.030	−0.377	−0.227
Recast	Pretest	−0.329	0.027	−0.402	−0.203
	Posttest	−0.320	0.031	−0.370	−0.193

Table 7.22

Linear mixed-effects model estimates for proportional difference in pitch by feedback provision

Formula: lmer (pitch ~ feedback * time + (1+ time | learner) + (1 + time | word))

	β estimate	*SE*	*df*	*t*-value	*p*-value
Intercept	−0.328	0.004	34.60	−92.555	0.000
FeedbackNoRecast	0.007	0.003	542.20	2.407	0.016
TimePosttest	0.009	0.002	116.60	3.643	0.000
FeedbackNoRecast: TimePosttest	−0.007	0.004	549.70	−1.702	0.089

Goodness of fit: $R^2 = 0.465$

Table 7.23

Predicted pretest and posttest means for proportional difference in pitch by feedback provision

Feedback	Time	β estimate	*SE*	*df*	lower *CL*	upper *CL*
No Recast	Pretest	−0.321	0.004	48.70	−0.329	−0.313
	Posttest	−0.318	0.005	56.63	−0.328	−0.309
Recast	Pretest	−0.328	0.004	34.60	−0.335	−0.321
	Posttest	−0.319	0.004	44.13	−0.327	−0.310

Table 7.24

Pairwise pretest-posttest contrasts for proportional difference in pitch by feedback provision

Feedback	β estimate	*SE*	*df*	*t*-value	*p*-value
No Recast	0.003	0.003	254.84	0.770	0.442
Recast	0.009	0.002	116.84	3.643	0.000

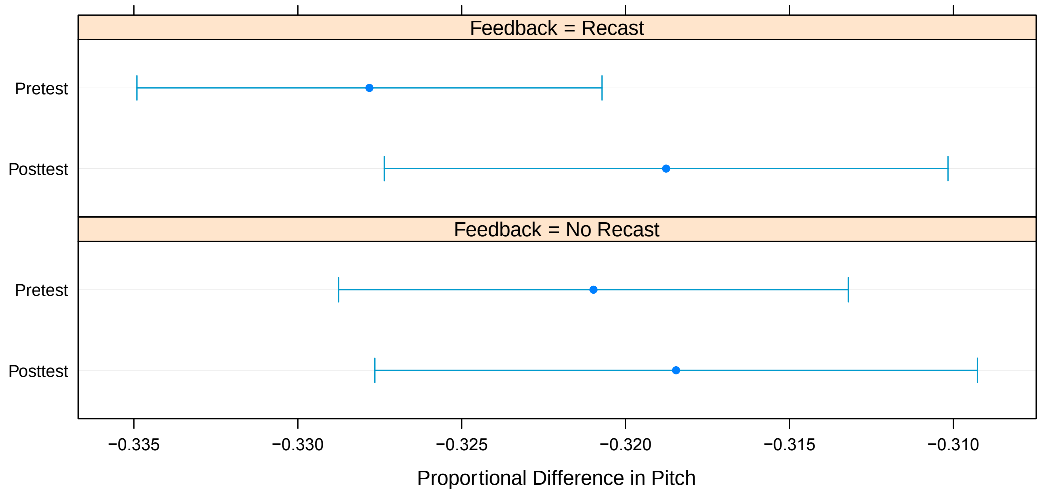


Figure 7.23 95% confidence intervals for proportional difference in pitch by feedback provision

According to the results, the intervention group produced the second syllables of the words that did not receive a recast with 0.3% higher pitch on the posttest than they did on the pretest. On the other hand, they produced the second syllable of the words that received a recast with 0.9% higher pitch on the posttest in comparison to the pretest. Further inspection of pairwise contrasts revealed that the difference was statistically significant for the words that received a recast, whereas the difference over time was not statistically significant for the words that did not receive a recast. Table 7.24 and Figure 7.23 show the pairwise contrasts.

The estimated distributions of pitch measures for both groups of words were also compared to the native-speaker baseline (Figure 7.24). The boxplots showed that although the pretest and posttest mean values of the words that did not receive a recast were comparable, these words were produced with a slightly wider range of pitch values on the posttest. The mean value of the words that received a recast moved closer to the mean value of the native-speaker baseline on the posttest. That said, the native-speaker data was considerably more spread out, capturing the pretest and posttest distributions of not only the words that received a recast but also the ones that did not. As a result, the majority of data points for both groups of words overlapped with the portion of native-speaker distributions between the first and third quartiles. The shape and amount of overlap was similar to the pattern that emerged earlier in the group-level analysis of pitch measures (Figure 7.18).

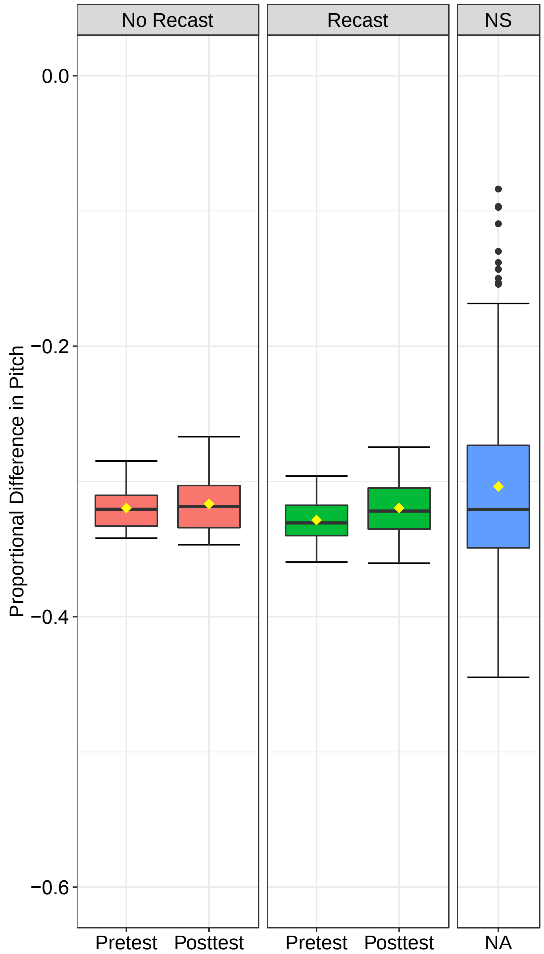


Figure 7.24 Comparison between NS and NNS productions of proportional difference in pitch by feedback provision

7.3. Summary of acoustic analyses

Two sets of acoustic analyses were carried out in order to analyze the impact of recasts on learners' production of syllable duration, intensity, and pitch, which are the main correlates of lexical stress in English. The first set of acoustic analyses, referred to as Acoustic Analysis 1, focused on the changes in the proportional difference measures over time by comparing the pretest and the posttest productions of learners in the two experimental conditions. The results of the first set of analyses showed that the intervention group produced the second syllables of the target words with statistically longer duration and higher pitch on the posttest. However, it needs to be mentioned that the degree of change was not the same for duration and pitch measures. The change over time in proportional difference in duration was 3.6%, which was larger than the gains in proportional difference in pitch, which reached only 0.7%. What is more, the model fitted for the analysis of duration had an R^2 value of 0.546, whereas the model fitted for the analysis of pitch had an R^2 value of 0.467. This means that the mixed-effects model used for the analysis of duration explained more of the variance in learner productions, therefore, had a larger effect size compared to the model fitted for the analysis of pitch. Learners' production of second syllable intensity increased by only 0.2%, which was not statistically significant. When it comes to the control group, the posttest productions of duration, intensity, and pitch measures did not yield statistically significant gains.

The learner-level and word-level visual exploration of the predicted values provided further insights into the differences between the two experimental conditions. The distribution of the predicted values showed that each learner in the intervention group produced the second syllables of the target vocabulary with longer duration and higher pitch on the posttest. That said, the gains in duration were more pronounced, and considering that a higher amount of the variance was explained by the mixed-effects model for duration, the learner-level data indicates that recasts had a positive impact on the realization of syllable duration. Once again, the control group data displayed no pretest-posttest contrasts at the learner or the word level. A large majority of the learners in the control group produced the three acoustic correlates of stress with a highly similar distribution on the pretest and the posttest. As each target word was produced by both experimental groups, it was also possible to make a direct between-groups comparison by inspecting the realizations of each target word. The word-level analysis showed that the second syllable of each and every single target word was realized with relatively longer duration and higher pitch on the posttest by the intervention group. Once again, the data was consistent at the word-level in that the differences over time were more pronounced in the case of duration measures than pitch measures as there was much less of an overlap between pretest and posttest distributions of duration. On the other hand, the pattern that emerged among the words produced by the control group remained stagnant over time. Almost every single target word produced by the learners in the control group had a similar distribution of duration and pitch measures on the pretest and the posttest indicating that no changes occurred over time. The replicability of the results at learner-level and word-level analyses provided further support for the consistency of the model estimates, the gains by the intervention group, and the effect of recasts, particularly on syllable duration.

Finally, the boxplot comparisons between learner and native-speaker productions of the three acoustic correlates showed that the posttest realizations of syllable duration and pitch by the intervention group moved in the direction of the native-speaker baseline. However, the magnitude of change was larger for the duration measures. The gains on the posttest were minimal for pitch measures. What is more, the pretest and posttest distributions of pitch by the intervention group both fell within the range of native speaker productions of pitch. In contrast, as for duration measures, the gap between the native-speaker baseline and the intervention groups' pretest productions was much larger, which reduced on the posttest. This could mean that there was more room for positive gains in the case of duration than there was in the case of pitch.

In short, the most notable finding from the first set of analyses is that the intervention group produced the second syllables of the target words with relatively higher duration, which was the case for every single learner in the intervention group and every single target word produced by the intervention group. Furthermore, the intervention group's posttest realizations of syllable duration moved in the direction of the native-speaker baseline. Although similar patterns were observed for the pitch values produced by the intervention group, the gains were not as large or as consistent as the gains observed in the case of duration. Considering the overall results of the first set of acoustic analyses, it was clear that for the learners who took part in the current study, recasts had the greatest impact on syllable duration among the three acoustic correlates of lexical stress.

As a post-hoc analysis, the correlation between word frequency and gains in duration were investigated. The purpose of this additional analysis was to understand whether word frequency facilitated or hindered gains in duration, the acoustic correlate that learners in the current study have improved the most upon receiving recasts. The word frequencies were retrieved from the Corpus of Contemporary American English (Davies, 2008-). The frequencies used in the analysis were based on the per-million values. A visual representation of the relationship between word frequency and gains in duration for the 20 words that were produced by the intervention group is shown in Figure 7.25. As can be seen in the figure, there was a curvilinear relationship between word frequency and gains in duration. Also, due to the fact that there were only 20 words, one word with the highest frequency skewed the correlation. In order to test the strength of the correlation, a Spearman's rank correlation test was run. The results showed that there was a weak correlation that was not significant (rho = 0.33, p = .15). It is difficult to reach a strong conclusion about the impact of word frequency on the gains that corrective feedback triggers as the post-hoc analysis is based on a small set of vocabulary. However, it was clear that in the current study, word frequency did not have a strong correlation with gains in duration.

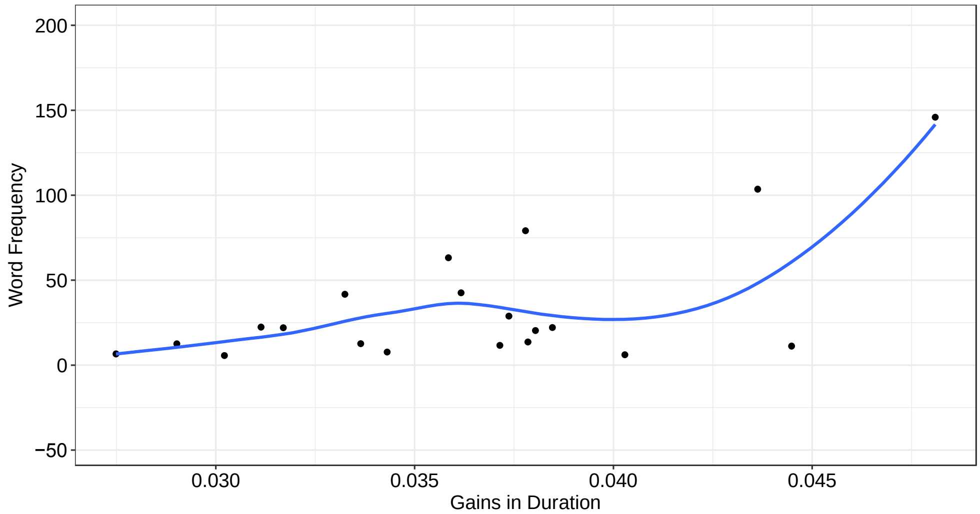


Figure 7.25 Correlation between word frequency and gains in duration

The second set of acoustic analyses, referred to as Acoustic Analysis 2, focused on the changes in the acoustic measures over time by comparing the words that received a recast and the words that did not. According to the results, the second syllables of the words that received a recast were produced with statistically longer duration and higher pitch on the posttest, whereas no significant differences were observed between the pretest and posttest productions of intensity. The posttest gains in duration was 5.5% for words that receive a recast and −0.2% for the words that did not receive a recast. Considering the fact that the intervention group exhibited an overall increase of 3.6% in syllable duration as it was earlier in Chapter 7.1.1, the second set of analyses indicated that the improvement was mainly driven by the words that received a recast. When it comes to the pitch measures, the words that received a recast were produced with 0.9% higher pitch on the posttest while the words that did not receive a recast were produced with 0.3% higher pitch. The gains in pitch measures were marginal in comparison to the gains in duration. Similar to the findings from the first set of acoustic analyses, the gains were more pronounced for syllable duration than they were for syllable pitch. These findings were supported further by the effect sizes for the mixed-effects models fitted for the analysis of duration (R^2 = 0.539) and the analysis of pitch (R^2 = 0.465). Finally, the analysis of the words that did not receive a recast showed that there were no statistical differences between the pretest and posttest productions of syllable duration, intensity, or pitch.

Finally, the boxplot distributions that compared learner productions to the native-speaker baseline were highly similar to the boxplot distributions created for the comparison between acoustic measures of words by feedback provision and the native-speaker baseline. The posttest realization of syllable duration for the words that received a recast became more aligned with the native-speaker baseline. Although recasts have also triggered a positive shift in the production of syllable pitch, the native-speaker baseline was more spread out, capturing the pretest and the posttest distributions of both groups of words produced by the intervention group. In this sense, the boxplots displaying the distributions of pitch measures in Acoustic Analysis 2 were very similar to the boxplots that displayed the distribution of pitch measures in Acoustic Analysis 1. Last but not least, once again there were no changes in the production of intensity over time.

The patterns that emerged in the second set of analyses provided further insights into the immediate impact that recasts have on the acoustic correlates of stress. Altogether, the results showed that out of all three acoustic correlates, the greatest amount of gains occurred in the case of syllable duration produced by the intervention group. Also, out of all the words that the intervention group produced, the words that received a recast were produced with relatively longer second syllable on

Table 7.25

Comparison of gains by group and by feedback provision

Acoustic Analysis 1					
	R^2	Intervention Group		Control Group	
		gain	*p*-value	gain	*p*-value
Duration	0.546	3.5%	.000	−0.1%	.950
Intensity	0.450	0.2%	.294	0%	.989
Pitch	0.467	0.7%	.001	0.2%	.353
Acoustic Analysis 2					
	R^2	Recast		No Recast	
		gain	*p*-value	gain	*p*-value
Duration	0.539	5.5%	.000	−0.2%	.909
Intensity	0.441	0.2%	.556	0.3%	.422
Pitch	0.465	0.9%	.000	0.3%	.442

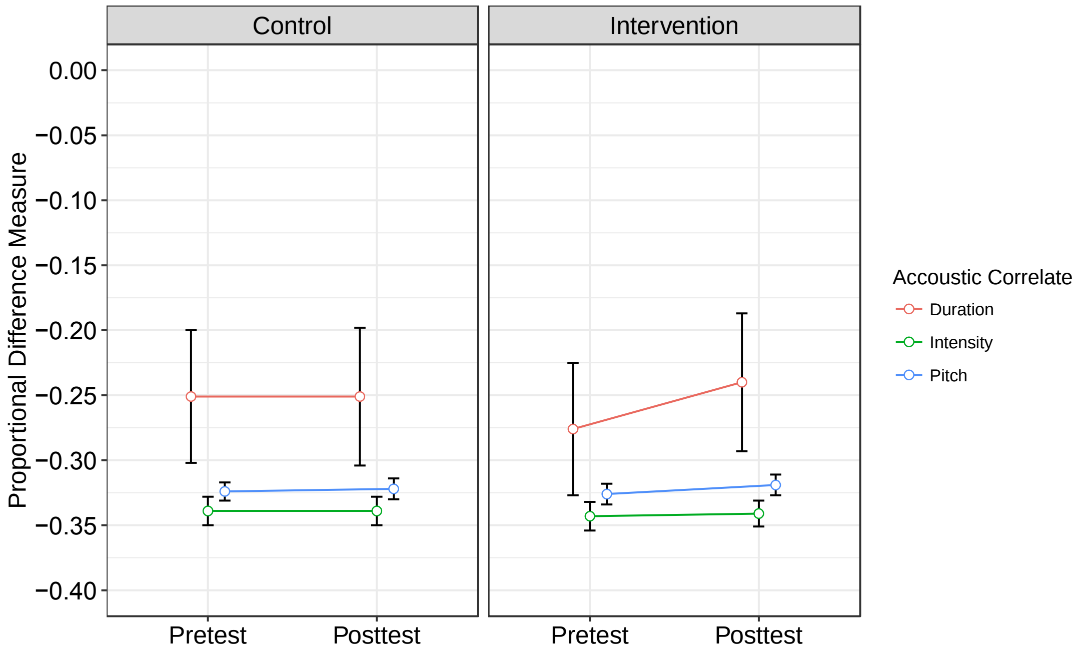


Figure 7.26 Confidence intervals for Acoustic Analysis 1

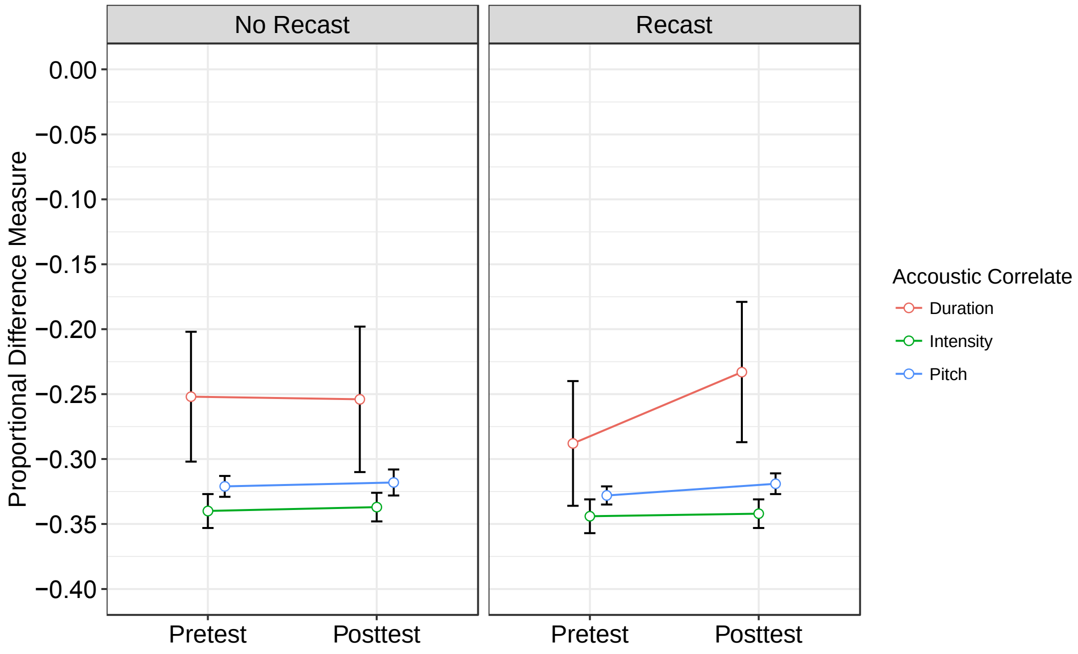


Figure 7.27 Confidence intervals for Acoustic Analysis 2

the posttest. Table 7.25 provides a summary of the gains and Figures 7.26 and 7.27 provide a graphical summary of confidence intervals for both sets of analyses.

Before concluding this section, it is probably necessary to briefly touch upon the differences in variation when listener and learner productions of the three acoustic correlates are compared as shown earlier by Figures 7.6, 7.12, and 7.18. The difference in variation between the two groups was considerably larger in the case of intensity and pitch measures with learner productions having much less variation compared to native speaker productions. Although there was also a difference in variation between learners' and native speakers' production of duration, the difference was not as stark as the case of pitch and intensity. There can be a few possible explanations for the observed patterns. To begin with duration measurements, the higher variation observed in learners' duration data compared to their intensity and pitch data could be due to the focus that learners placed on duration. Learners in the current study attempted to utilize duration more than intensity and pitch to shift their stress placement confirming findings from previous research which argued that L1 Arabic speakers mainly rely on syllable duration when assigning primary stress in L2 English (Munro, 1993; Parlak & Ziegler, 2017). As learners manipulated duration more than the two other acoustic correlates in an attempt to improve their primary stress placement, it is logical to think that they experimented with syllable duration in various ways which may have contributed to the higher variation observed in their production of duration, which looked more similar to the variation in native speakers' production of duration. In the same way, one of the reasons why learners' production of intensity and pitch varied less could be that these two acoustic correlates did not receive the same level of attention, and therefore, they were manipulated to a lesser degree resulting in less variation.

Another contributing factor could be the differences between learners' native language and the target language in terms of their respective vowel inventory. The vowel inventory of Arabic is considerably limited compared to that of English. Also, Arabic has very few complex vowels such as diphthongs which are only realized in limited contexts regardless of the dialect (Mustafawi, 2018). In addition to having a limited vowel inventory at their disposal, L1 Arabic speakers do not reduce unstressed vowels to schwa. The schwa usually has less intensity compared to a full vowel and it does not attract a high pitch. Therefore, when L1 Arabic learners do not reduce unstressed vowels to a schwa and produce all syllables with a full vowel, and because they have a limited number of vowels in their phonological inventory, it would not be surprising that there is less variation in their production intensity and pitch. In comparison, native speakers of English have a larger vowel inventory and they reduce unstressed vowels to a schwa possibly leading to more variation in their production of intensity and pitch.

In addition to the differences in phonological inventory, another reason why learners had relatively less varied pitch productions could be related to performing in a foreign language. Previous research has shown that L2 speakers have lower pitch ranges in the target language compared to the pitch ranges produced by the native speakers of the target language. Zimmerer, Jügler, Andreeva, Möbius, and Trouvain (2014) investigated L1 German speakers' pitch ranges in L2 French as well as L1 French speakers' pitch ranges in L2 German. Their results showed that both L1 German and L1 French speakers produced smaller pitch ranges in their L2. The authors offered two possible reasons for smaller pitch ranges in L2. As one possible reason, they speculated that speakers were probably not confident about the location and realization of correct pitch values in their L2. The second possible reason they offered was that speakers may have focused more on the correct production of segmental features which may have taken their attention away from realizing suprasegmental features correctly. The authors also looked in to the effect of proficiency level, and although their results were inconclusive, they argued for a positive correlation between proficiency level and pitch ranges. In another study that focused on L1 Italian speakers' production of pitch in L2 English, Busà and Urbani (2011) found that Italian speakers' pitch production had less variation compared to the variation in pitch produced by native speakers of English. They suggested that the differences may be due to socio-cultural or cross-linguistic influence. Narrower pitch ranges in L2 were also observed in other studies that investigated L1 Dutch speakers' pitch ranges in L2 Greek (Mennen, 1998) and L1 Finnish speakers' pitch ranges in L2 Russian (Ullakonoja, 2007). Although these studies looked into pitch ranges at the phrase level, similar effects would naturally be expected at the word and syllable level. In short, learners in the current study may have produced less varied pitch ranges compared to native speakers due to a combination of the potential reasons raised by these studies.

7.4. Listener judgments

The interrater reliability for listeners' judgment of primary stress placement was checked using the irr package version 0.84 (Gamer, Lemon, Fellows, & Singh, 2012). Fleiss' κ was run to determine the level of interrater reliability among the 10 native-speaker listeners who judged the placement of primary stress by listening to the 120 three-syllable words that were prepared for the listener judgment experiment. The results showed that the level of agreement among the raters was very poor ($\kappa = .068$, $z = 6.55$, $p = .000$). For diagnostic purposes, a series of unweighted Cohen's κ tests was run for pair-wise analyses. As can be seen in Table 7.26, no two raters reached a desirable level of agreement (equal to or larger than .80).

The listener judgement data was also manually examined in order to identify potentially interesting patterns. Despite the overall disagreement among listeners, they unanimously agreed that the intervention group produced only 1 word out of 30 words with stress on the second syllable on the pretest while this number went up to 6 out of 30 words on the posttest. As for the judgment of words produced by the control group,

Table 7.26

Cohen's Kappa analysis of stress placement judgments

	NS 1	NS 2	NS 3	NS 4	NS 5	NS 6	NS 7	NS 8	NS 9
NS 2	0.140								
NS 3	0.133	0.122							
NS 4	−0.013	0.132	0.082						
NS 5	0.072	0.086	0.117	0.083					
NS 6	0.029	0.036	0.028	0.000	0.176				
NS 7	0.082	0.094	0.202	0.062	0.336	0.091			
NS 8	0.229	0.054	0.043	0.081	0.052	0.030	0.128		
NS 9	0.221	0.005	0.088	0.013	0.147	0.085	0.094	0.298	
NS 10	0.039	0.050	0.043	0.015	0.091	0.663	0.111	0.024	0.043

the listeners unanimously agreed that the control group produced 2 words out of 30 words with stress on the second syllable on the pretest and only 1 word out of 30 words on the posttest was judged to carry primary stress on the second syllable.

7.5. Exit survey

In addition to the acoustic analyses and listener judgments, the responses that learners provided on the exit survey were analyzed. The exit survey contained questions that were aimed at eliciting responses that provide evidence for noticing the corrective force of recasts (see Appendix D). As such, some of the questions directly addressed pronunciation related issues and others were used as a distractor (e.g., addressing vocabulary and grammar). In this section, only responses to questions that addressed pronunciation issues are reported.

There were two likert-scale items that asked learners to indicate how much they have focused on pronunciation during the interview task and whether they received feedback on pronunciation issues. Learner responses to the first item showed that in the case of the intervention group, 31% strongly agreed and 36% percent

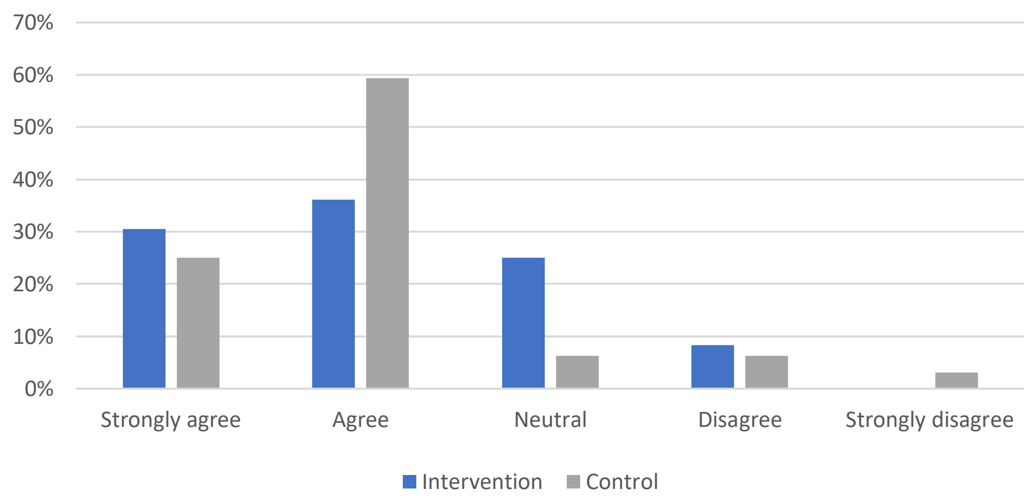


Figure 7.28 Learners' responses to the item "During the interview task, I focused on pronunciation" by group

agreed with the statement, whereas among the control group learners 25% of them strongly agreed and 59% of them agreed with the statement (Figure 7.28). When the responses for "strongly agree" and "agree" are added together, they come to 67% for the intervention group and 84% for the control group. That means, based on this particular survey item, the control group learners seem to have paid more attention to their pronunciation during the interview task. There could be various reasons for why the control group learners said they paid more attention to pronunciation such as being more musical, perceiving pronunciation as a weakness that needs to be worked on, or even the value placed on pronunciation as a language skill. In any case, it was clear that the attention to pronunciation was self-directed in the case of the control group learners as they did not receive any form of corrective feedback. In fact, learners' responses to another likert-scale item, which specifically targeted noticing of corrective feedback, supports this interpretation. This other likert-scale item asked learners whether the researcher provided feedback on their pronunciation errors. The difference between the two groups was striking in terms of the percentage of learners that strongly agreed with the survey item, which was 42% for the intervention group

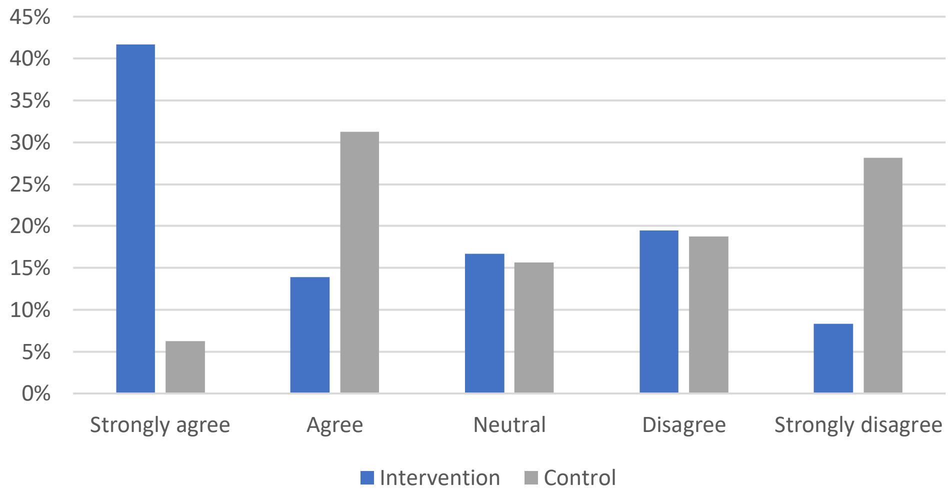


Figure 7.29 Learners' responses to the item "The interviewee gave me feedback when I made pronunciation mistakes" by group

and 6% for the control group (Figure 7.29). When the responses for "strongly agree" and "agree" were added together, it was seen that 56% of the intervention group and 38% of the control group agreed with this particular survey item. Interestingly, the gap between intervention and control group learners became wider when they responded to a yes/no question that required further elaboration in the form of an open-ended response. When they were asked "Did the interviewee correct your English during the interview task? If yes, what did he correct and how did you react? Please explain.", 74% of the learners in the intervention group responded to the first part of the item with a yes, whereas the number was only 3% for the learners in the control group (Figure 7.30). When it comes to the open-ended part of the item that required learners to elaborate on the type of correction, there were only a few responses provided by the control group learners which were related to the questions about word meaning or the task itself that they had raised at the onset of data collection prior to the interview activity. Following are some of the responses by the control group, each quote representing a different learner: "*in some words yes"; "yes, it was an indirect way when I asked him the questions that was related to it"; "yes, he corrected one of the*

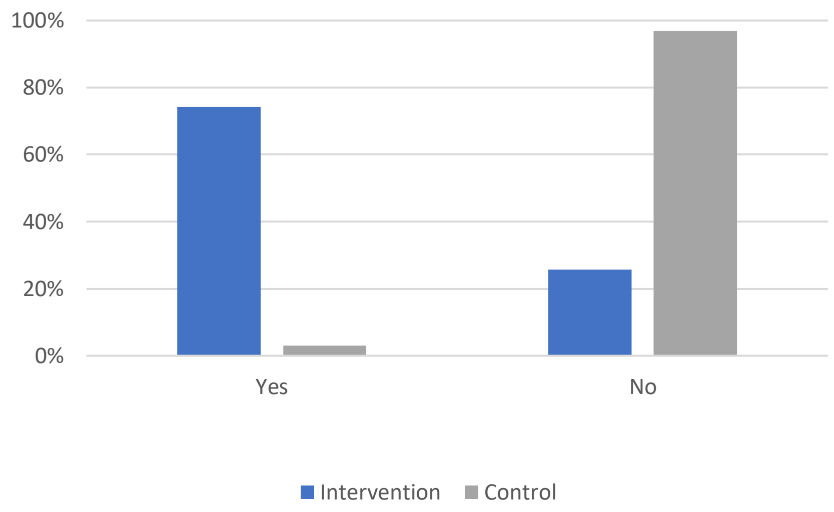


Figure 7.30 Learners' responses to the item "Did the interviewee correct your English during the interview task?" by group

words that I spelled"; "yes, consensus. His response was really good. He didn't get mad or shout. He was quite and easy."

In contrast, the intervention group learners provided more specific and detailed comments that provided evidence for identification of the focus of correction as pronunciation and in many cases accompanied by examples from the target lexical items. It was also clear from the responses given by the intervention group that they were referring to the feedback they received during the interview task. Some of the responses from the intervention group are "*he corrected my pronunciations"; "yes, he corrected some words I pronounced wrong"; "yes, he correct for me the in correct word like perception"; "yes, when I say consensus wrong, he told me the right way to say it and I say it correctly"; "dynamic"; "yes, the word contention"*. What is more, some of the responses provided by the intervention group provided further details about the nature of correction by identifying repetition of the word as a feature of the feedback they received, which indicates that the intervention group was able to notice the corrective force of recasts: *"pronunciation, he repite it"; "yes, he corrected my pronunciation by repeating the same word again"; "he also repeat some word more than one time"; "Yes, when I asked him the questions he used to repeat the key word which is the word that I don't know how to read it correctly"*. As can be seen from these examples, the responses given by the intervention group were in stark contrast to the responses given by the control group.

There was one more open-ended question that aimed to elicit learner responses that would provide clues about whether they noticed the recasts they received or not: *How much attention (if any) do you feel you paid to pronunciation during the interview task? Can you give any examples?* The responses provided by the control group clearly showed that many of the learners did not particularly focus on pronunciation, and when they did, the attention they paid to pronunciation was self-directed rather than feedback-driven. Some of the responses were *"not that much"; "i paid little attention"; "no attention"; "I didn't really paid attention to my pronunciation"; "I paid attention to his pronunciation"; "I really tried to pronounce the words correctly";* and *"I give a little bit attention to the pronunciation and the way that i was pronouncing the words".*

In contrast, the intervention group answered the question in a way that showed evidence of noticing the corrective feedback that they received. In many cases, the intervention group learners referred to the target lexical items. Following are some of their responses with the target vocabulary underlined: *"i paid attention about how he pronunciate the word so if i said a word wrong now i know from him how to say it like contention"; "1) convention, 2) compassion, 3) consensus"; "A lot. The word contention and consensus"; "a lot of attention, like consistent"; "i paid attention to consensus and companion"; "100% consensus word"; "yes, the word perception"; "I gave a high attention for some difficult words such as suspicion"; "i was trying to pronounce the words correctly for example (compassion , consensus)".*

There were also responses indicating that learners paid attention to the extra emphasis used to enhance the salience of recasts such as *"a lot of attention, i liked the way he pronounced the words and explained his points of view using different tones to clarify the most important points,"* as well as responses that showed they noticed the recast that their partner received such as *"a lot when he correct to my partner."* In summary, the responses to the exit survey provided evidence for noticing of the corrective force of recasts by the intervention group.

CHAPTER 8: DISCUSSION AND CONCLUSION

SLA studies that adopt the interactionist approach have demonstrated the effectiveness of recasts in promoting language development in various contexts, for different age groups as well as target language features. Despite the established role of recasts as a corrective feedback move that promotes language learning (Goo & Mackey, 2013; Mackey et al., 2003; Mackey & Philp, 1998; McDonough & Mackey, 2006; Loewen & Philp, 2006; Philp, 2003), the relationship between recasts and L2 phonological development is still less clear due to a lack of interactionist studies that focus on L2 phonology. Recently, Saito and Lyster conducted a series of studies with L1 Japanese learners of English showing that recasts facilitate development of the approximant /ɹ/ (Saito 2013a, 2015a; Saito & Lyster, 2012a). The findings from their studies are promising as they show that research endeavors channeled toward this under-explored area have the potential to yield useful findings. In this respect, the current study was an attempt to contribute to the emerging group of studies that investigate the relationship between corrective feedback and development of L2 phonology by focusing on the impact of recasts on the development of lexical stress. By investigating the development of a suprasegmental feature, the current study was also an attempt at studying an under-explored feature within an under-explored area. This dissertation study provides evidence that recasts have the potential to trigger a positive shift in acoustic realization of primary stress after a brief intervention period. As such, the findings provide further support for the potential usefulness of recasts when treating L2 learners' pronunciation errors. The findings also fit in the wider SLA literature that has shown the efficacy of recasts when addressing the development of various morphological and syntactic features. The implications gained from the current study, which are discussed in more detail in the next sections, will hopefully provide directions for future research studies that aim to investigate the relationship between corrective feedback and L2 phonological development.

8.1. Summary of answers to research questions

The first research question asked whether providing recasts on learners' primary stress errors during a communicative activity is associated with more target-like production of primary stress. The second and third research questions directly stemmed from the first question in order to investigate further details on the potential relationship between recasts and primary stress placement. The second research question asked how recasts impact learners' acoustic realization of primary stress, while the third research question asked how recasts impact listener judgments of learners' stress placement.

To begin with the first research question, two sets of acoustic analyses carried out for the current study showed that there is a direct relationship between recasts and more target-like production of primary stress. The current study showed that learners who received recasts on their primary stress errors produced target words with statistically longer duration and higher pitch on the posttest with larger differences as well as larger effects observed in the case of duration. The analyses also showed that the effect was consistent across learners in the intervention group as well as across all target words produced by the intervention group. Based on these findings, the current study suggests that recasts can play a facilitative role in helping learners move toward target-like placement of primary stress.

When it comes to the second research question, the acoustic analyses yielded a clear pattern showing that learners in the intervention group produced the second syllables of the target words with longer duration in their attempt to shift primary stress placement onto that syllable. The intervention group also produced the second syllables of the target words with higher pitch on the posttest; however, the gains were

Table 8.1

Summary of answers to research questions

Research Question	Data Used to Answer the Research Question	Summary of Analyses	Answer to the Research Question
RQ1: Is providing recasts on learners' primary stress errors during a communicative activity associated with more target-like production of primary stress?	Data 1: Duration, intensity, and pitch measures from target words produced by learners and native speakers Data 2: Native-speaker judgments of learners' primary stress placement on the target words	Data 1: Learners' pretest and posttest productions were compared to the native-speaker baseline, which showed that duration and pitch measures became more target-like on the posttest. Data 2: The results from listener judgments were inconclusive.	Acoustically, learners' productions became more target-like on the posttest. More research is necessary to understand the impact of recasts on listener judgments.
RQ2: How do recasts impact learners' acoustic realization of primary stress?	Duration, intensity, and pitch measures from target words produced by learners	Linear-mixed effects analyses indicated significant gains on the posttest for duration and pitch measures. No changes were observed in the case of intensity measures.	Learners mainly employed duration and to a lesser extent pitch when attempting to place primary stress correctly. The heavy reliance on duration could be due to cross-linguistic influence.
RQ3: How do recasts impact listener judgments of learners' stress placement?	Native-speaker judgments of primary stress placement	The analysis showed that it was difficult for listeners to indicate the location of primary stress in words produced by learners	More research is necessary to understand the impact on listener judgments.

not as large as the gains for duration. Finally, recasts did not lead to an increase in intensity on the posttest. Based on these findings, duration emerged as the acoustic correlate that was utilized the most by the learners in the current study upon receiving recasts on their primary stress errors.

As far as the final research question goes, the effects of recasts on listener judgments were inconclusive. The development of lexical stress triggered by recasts was most likely at an in-between stage making it difficult for listeners to decide which syllable carries the primary stress.

8.2. Discussion of findings

8.2.1. Changes in acoustic correlates

Before elaborating more on the findings based on the acoustic analyses, it is important to reiterate that target-like production of primary stress refers to production of the target syllable with longer duration, higher pitch and/or intensity. As it was mentioned earlier in Chapter 4.3, there is no phonetic formula that dictates in what combinations the three correlates come together to form what is perceived as primary stress; in addition, no single correlate is a predictor of lexical stress on its own (Cutler, 2005; Lehiste & Peterson, 1959; Lieberman, 1960). That said, the majority of lexical stress studies have shown that duration is a highly reliable predictor of stress (Gordon & Roettger, 2017). Also, duration is a cross-linguistic correlate that reliably predicts primary stress in a number of languages (Arvaniti, 2000; Dogil & Williams, 1999; Ortega-Llebaria, 2006; Sluijter et al., 1997) including Arabic (de Jong & Zawaydeh, 2002) and English (Turk & Sawusch, 1996). Gordon and Roettger's (2017) survey of phonetics research that focused on lexical stress also showed that duration has been identified as a robust predictor by a high percentage, 85% to be specific, of the studies that were surveyed. Although duration cannot be taken as a single predictor of stress, research has shown that it is the most consistent predictor across a number of languages which increases its importance when interpreting the results of the current study. The acoustic measures in the current study were calculated as proportional difference, also referred to as the Michelson contrast ratio by Taylor and Wales (1987) who argued that it is a highly robust ratio when measuring acoustic contrasts. An increase in the acoustic correlates of stress determined by the changes in proportional difference measures was taken as an indication of more target-like production of primary stress. Comparisons between the acoustic measures produced by the two experimental groups and the native-speaker baseline also provided information on the extent of gains that the intervention group was able to achieve.

Another important point to mention is that the acoustic realization of segmentals and suprasegmentals falls on a continuum in the acoustic space. To be more specific, there can be a range of possible acoustic values that determine the perception and production of a particular vowel and consonant. The same principle applies to the production and perception of syllable duration, intensity, and pitch leading to acoustic realization and perception of primary stress to also occur within a range of values. For example, Saito and Lyster (2012a) explain that native-speaker listeners' judgment of L1 Japanese speakers' production of /ɹ/ - /l/ contrast occurs on a continuum. They found that native-speaker listeners perceive learner productions as /ɹ/ when F3 values are around 2,270 Hz and as /l/ when F3 values are around 2,800 Hz. Sounds that fall between these values, particularly within the 2400-2600 Hz range, were perceived as either /ɹ/ or /l/. Similarly, in a study that investigated the acoustic realization and perception of lexical stress in disyllabic words produced by L1 Mandarin speakers, Lai, Sereno, and Jongman (2008) showed that acoustic correlates of stress calculated as first-to-second vowel ratios were realized as a range of values by native speakers, beginning learners, and advanced learners. The advanced learners' productions were closer to the native speaker range for all the correlates, whereas the range of realizations by the beginning learners was more distant from the range of native-speaker realizations. As can be seen from these examples, what counts as target-like in L2 phonology studies can be different from what may count as target-like in studies that focus on development of morphology or syntax. In the case of morphology and syntax, identifying development can be relatively straightforward. For example, a researcher may measure development of English past tense by analyzing the percentage of correct usage in obligatory contexts. As usage of past tense would be labelled as either correct or incorrect, at its core the analysis would be based on a dichotomy rather than a range of values on a continuum. On the other hand, when measuring L2 phonological development, researchers need to consider the degree of change in the acoustic correlates that constitute the target feature as well as the direction of that change. Based on this brief discussion, in the current study, the term target-like is used to refer to posttest production of an acoustic correlate of lexical stress that is statistically different from pretest productions and has changed in the direction of the native-speaker base line.

The current study showed that L1 Arabic speakers relied on duration more than pitch or intensity when trying to improve their primary stress placement in English. This finding has implications in terms of cross-linguistic influence as it provides further support for previous research as well as the pilot study which showed that L1 Arabic speakers rely heavily on syllable duration when producing lexical stress in English. Previous research by Flege and Port (1981) highlighted cross-linguistic influence on English vowels produced by L1 Arabic speakers. They found that there were no differences between the duration of English vowels preceding voiced and voiceless stops when produced by L1 Arabic speaking participants. In contrast, native speakers of English produce vowels preceding a voiced stop with longer duration. In a different study, Munro (1993) noted that Arabic speakers may perceive tense-lax vowel contrasts in English as a matter of duration rather than vowel quality and attempt to produce these contrasts by utilizing vowel duration as they do in their native language to differentiate between short and long vowels. In short, duration as a crucial aspect of the phonology of Arabic plays an important role when L1 Arabic speakers produce vowels in English.

The pilot study for the current PhD dissertation also showed that L1 Arabic speakers utilize duration more than other correlates of stress when attempting to shift primary stress onto the correct syllable (Parlak & Ziegler, 2017). The parallelism between the findings of the pilot study and the current study is important in that duration emerged as the main acoustic correlate of stress that L1 Arabic learners tried to utilize when improving their primary stress placement. As both the pilot and the current study were conducted with a relatively large number of participants within the context of interactionist SLA research, 57 learners in the pilot study and 64 learners in the current study, the findings indicating the crucial role of duration were consistently robust. This finding is also in line with studies that investigated lexical stress in Arabic and established a consistent link between stress and duration (e.g., de Jong & Zawaydeh, 1999). In other words, a highly possible explanation for why the learners in the current study used duration in an attempt to shift stress onto the second syllable would be that their native language, Arabic, has "a very robust and systematic vocalic contrast which depends on vowel duration" (de Jong & Zawaydeh, 2002, p. 71). This is most likely the reason why learners in the current study did not utilize pitch and intensity as much as duration when attempting to move stress onto the second syllable. There were virtually no pretest-posttest differences in learners' production of proportional intensity regardless of experimental condition. As for pitch, although learners in the intervention group showed gains on the posttest, the increase in pitch measures was less than the increase in duration. Another important observation is that the production of syllable pitch was already within range of the native speaker baseline with similar mean and median values as shown in the boxplots that compared learner productions to the native-speaker baseline. Therefore, the increase in learners' production of pitch was not as consequential as the increase in duration measures, which displayed a larger difference from the native speaker baseline on the pretest and moved closer to the it on the posttest.

8.2.2. The benefits of recasts for the development of lexical stress

Based on the results of the acoustic analyses, the current study suggests a positive link between recasts and development of primary stress placement. The first set of acoustic analyses showed that recasts led the intervention group learners to produce the second syllable of the target words with statistically longer duration and higher pitch on the posttest which was consistent across all intervention group learners and all target words produced by them. The second set of acoustic analyses revealed further support for the role of recasts by showing that the words that received a recast were produced with statistically longer duration and higher pitch, whereas there were no statistical changes over time for the words that did not receive a recast. This finding suggests that learners were able to notice the corrective force of recasts and attempted to change their production of the target words that received a recast but not the ones that did not receive a recast. These results have a number of implications in terms of the role that recasts play in promoting language development. As Long (2007) suggests, the contingency of recasts which provides opportunities for juxtaposition between the target and the non-target-like production is what makes recasts useful regardless of what the target language feature is. There has been substantial amount of evidence supporting Long’s view and showing that recasts are useful for development of morphological and syntactic features (Goo & Mackey, 2013; Li, 2010, 2014). Recently, interactionist studies have shown that recasts can also be useful for the development of segmental contrasts (Saito, 2013a, 2015a; Saito & Lyster, 2012a). The current study provides further support for the role of recasts by showing a positive relationship between recasts and development of lexical stress.

There are a number of potential reasons why recasts were able to trigger a positive change in learners' acoustic realizations of the target words.

SLA literature suggests that recasts are particularly effective when addressing features that are crucial in communicative value (Carpenter et al., 2006; Long, 2007). Therefore, one explanation for the impact of recasts on the development lexical stress observed in the current study would be that phonological errors are highly crucial in terms of communicative value. As mentioned in Chapter 4.4, lexical stress cues carry vital acoustic information that has an impact on listeners' online word recognition (e.g., Cooper et al., 2002; van Donselaar et al., 2005). Listeners rely on lexical stress cues when retrieving vocabulary from their mental lexicon. Research has also shown that listeners rely on phonological information, which includes lexical stress, when judging comprehensibility and accentedness of L2 speech (Saito et al., 2016, 2017). Although comprehensibility of L2 speech is affected by various lexical, grammatical, and phonological factors, lexical stress is a major one among those factors in terms of its impact on comprehensibility (Saito et al., 2017). As such, lexical stress errors are crucial in terms of how much they hinder communication. Although the recasts used in the current study were didactic recasts and no real communication breakdown occurred due to the fact that the researcher knew which target words the learners were saying, the interview task was a simulation of a natural conversation and the recasts targeted primary stress issues that are inherently important when it comes to communicative value.

Another factor that potentially contributed to the effectiveness of recasts in the current study was their psychophysical salience (N. Ellis, 2016, 2018). The recasts in the current study were presented in isolation as a single word with emphatic stress on the correct syllable and then repeated as part of the response given to learners. This particular procedure followed for feedback provision allowed recasts to stand out in the input and most likely facilitated noticing by the learners. What is more, the fact that the recasts in the current study were short and had a single change may have helped with the working memory load (Philp, 2003) and allowed learners to notice the corrective force of recasts and juxtapose it with their immediately preceding production. When recasts are short and isolated, it becomes easier for learners to notice them, store them in their working memory for a brief period of time in order for the juxtaposition between the non-target-like and target-like forms to occur. Furthermore, the same characteristics that made recasts in this study more salient also made them more explicit. As SLA researchers have argued in the past, recasts that are more explicit are potentially beneficial for learners (Ellis et al., 2006; Ellis & Sheen, 2006; Sheen & Ellis, 2011). Therefore, the explicit nature and the increased psychophysical salience of the recasts used in the current study most probably facilitated noticing by learners and allowed them to benefit from the feedback. Another important aspect that increases the effectiveness of recasts is their intensity (Mackey & Philp, 1998) and consistency (Han, 2002). When recasts focus on a particular language feature and are abundant in number, it becomes easier for learners to notice the correction and benefit from it. The recasts in the current study exclusively focused on lexical stress errors. What is more, many learners received a number of recasts, and in addition, they were exposed to the recasts that their partners received. The intensive provision of recasts may have also led to a priming effect and revealed the corrective force as well as target of recasts to learners. Correspondingly, there is evidence from the exit survey showing that noticing of the recasts by learners indeed took place. 74% of learners in the intervention group agreed that their English was corrected during the interview task, whereas only 3% for the control group learners thought that they received some form of correction. These numbers indicate that learners in the intervention group were able to notice the corrective force of recasts. However, for corrective feedback to work, learners not only need to understand the corrective force of the feedback but also the source or nature of the linguistic problem (Pinker, 1989; Roberts, 1995). The responses given to the open-ended items provided further evidence indicating that learners actually understood that the recasts they received targeted pronunciation issues. The majority of the responses specifically mentioned the target words in response to the questions that asked about the type of correction or whether they paid attention to pronunciation issues. Also, some of the responses specifically mentioned target words along with pronunciation-related comments such as “saying it wrong” or “pronouncing it wrong.” Yilmaz (2016) warns against retrospective verbal reports as they may not reveal information about noticing due to various issues such as memory constraints or learners’ inability to describe their experience. Considering the possibility that this could also be true in the case of the current study, the exit survey results become even more meaningful. That is to say, some of the learners in the intervention group who did not mention specific target words in response to the survey questions may have done so due to memory constraints. In the same way, learners who specifically mentioned target words in their responses did not mention every single target word but a few representative ones. This could also be due to the fact that they were not able to remember all the words for which they received a recast or perhaps they thought a few examples would be enough. Overall, the survey responses indicate that learners in the intervention group were able to notice recasts. In this way, the exit survey provides support for previous research which highlighted that feedback on phonological errors are easily noticed (e.g., Carpenter et al., 2006; Kim & Han, 2007; Lyster & Saito, 2010; Mackey et al., 2000) and acted upon (Ellis et al., 2001; Lyster, 1998a) by L2 learners.

8.2.3. Listener judgments

The results from the listener judgment experiment were inconclusive as it was difficult for listeners to identify the location of primary stress in the words that they heard. It is unlikely that this difficulty was due to the way the rating session was structured. The listening experiment required listeners to listen to target words in isolation and determine whether the first, the second, or the third syllable was stressed by indicating their choice on a nominal scale. Listeners were also given a chance to listen to a word multiple times before they made a judgment. Moreover, listener fatigue could not have been a problem either as the number of tokens and the length of the entire listening experiment were reasonable. There was only a total of 120 tokens that listeners rated and they were allowed to take a break in the middle of the rating session. The entire data collection session lasted for about 20-25 minutes. When it comes to their background, listeners shared a common background as they were all native speakers of North American English who were familiar with the variety of English spoken by L1 Arabic speakers. They also shared a highly similar socio-economic background which allowed them to travel the world, experience different cultures, and be exposed to different varieties of English spoken by a wide range of non-native speakers. Also, they all lived and worked in the United Arab Emirates, where they come in contact with non-native speakers from many different countries on a daily basis. Previous research has shown that raters' background can have an impact on their judgment of non-native speech. That said, familiarity with the L2 speakers' L1 or interlanguage is not necessarily a negative factor as it has been shown that familiarity with the non-native speakers' L1 in fact facilitates comprehension

(Gass & Varonis, 1984; Kennedy & Trofimovich, 2008). Research has also shown that previous experience with the L2 speaker's L1 can lead to positive bias when assessing their pronunciation (Carey, Mannell, & Dunn, 2011; Winke & Gass, 2013). These positive effects could be due to increased perceptual accuracy at the word and sentence level, which is a result of the familiarity with a particular non-native accent (Bradlow & Bent, 2008; Clarke & Garrett, 2004). Based on these research findings, raters' familiarity with the variety of English spoken by L1 Arabic speakers is not considered to be a negative influence on their judgment of stress placement. Although certain personal characteristics such as multilingualism and socio-biographical differences such as social environment or education level can also have an effect on listeners' judgment of non-native speech (Dewaele & McCloskey, 2015), the native-speaker raters who participated in the current study did not differ too much in terms of these factors either. One major point in which the listeners in the current study differed was that six of them were monolingual speakers of North American English and four of them spoke a heritage language in addition to speaking North American English. At first glance, this may seem like a possible reason that led to the inconclusive results. However, if that had been the case, then we would at least expect the ratings by the remaining six monolingual listeners to be aligned either as a group or in pairs. However, that did not happen either, and therefore, speaking a heritage language is possibly not the main reason for the lack of agreement among listeners.

It also needs to be noted that studies that have investigated rater bias or differences usually focus on judgments of comprehensibility, accentedness, or global scores assigned to pronunciation accuracy. These judgments are usually based on 5-point to 9-point likert scales or rubrics with various levels of definition. In other words, the rating instruments used in these studies are more complex in nature compared to the 3-point nominal scale used in the current study. The task of determining primary stress in a three-syllable word should have been more straight-forward than providing a subjective judgment of comprehensibility or accentedness. In light of these points, the difficulty experienced by native-speaker listeners during the rating session was probably caused by something else rather than factors related to rater background.

Upon completion of the rating session, listeners were asked whether they faced any difficulties when making judgments of stress placement. One point that they all mentioned was the unusual patterns in the acoustic correlates which made it difficult for them to pinpoint the location of primary stress. Listeners explained that some of the words sounded to have one syllable with longer duration, or more energy and emphasis, while another syllable with higher pitch. As a result, they were indecisive when identifying the location of stress for some of the words. The reason for the perceptual difficulties experienced by listeners is possibly the complex nature of the target feature and how its development occurs gradually on a continuum. In other words, it could be that recasts triggered a shift toward more target-like production in learners' interlanguage, which is supported by the acoustic analyses, but the shift was not large enough to allow listeners to confidently determine where the primary stress falls. The facilitative effects of recasts on syllable duration and pitch observed across learners as well as target words were robust, particularly in the case of duration. However, as the development of phonological features is not binary and occurs gradually, learners may need more opportunities for practice that would allow exposure to exemplars of the target form as well as more feedback on their non-target-like production in order to be able to further the positive shift triggered by recasts in their interlanguage. This would help learners move beyond the developmental limbo toward more target-like production, which would then make their it easier for listeners to perceive the location of primary stress in their output. Similar to the finding by Saito and Lyster (2012a) which showed that F3 values within the range of 2400-2600 Hz range were perceived as either /ɹ/ or /l/ by the raters, the productions of lexical stress by learners in the current study may have fallen into an in-between range making it difficult for listeners to decide syllable carries the primary stress.

Another debilitating factor could be that L1 Arabic speakers do not reduce the vowels in unstressed syllables to a schwa (Almbark et al., 2014; Zuraiq & Sereno, 2007). Once again, this is most likely due to a transfer effect from learners' first language. The vowel system of Arabic is a basic one with only three cardinal vowels /i, u, a/, and the entire vowel system of Modern Arabic comprises a total of six vowels, namely /i, iː, u, uː, a, aː/, that are formed through contrastive lengthening (Embarki, 2013). As vowel reduction is a not a phonological feature in Arabic, in general L1 Arabic speakers do not produce unstressed syllables with a schwa. However, when vowels in unstressed syllables are not reduced to schwa in L2 English, native speakers may experience comprehension difficulties. Braun, Lemhöfer, and Mani (2011) explored native English speakers' recognition of English words produced by L1 Dutch speakers. The author's acoustic analysis of Dutch-accented English and native English showed that suprasegmental cues for stress were comparable in both types of English; however, when unstressed syllables were analyzed, they observed differences in vowel quality as native English vowels were reduced to a schwa whereas Dutch-accented vowels were only slightly centralized. Their study also showed that the difference in vowel quality hampered native English listeners recognition of the Dutch-accented words. The authors concluded that although Dutch-accented English is generally intelligible, vowel quality in unstressed syllables produced by L1 Dutch speakers may hamper listener perception. A study by Cooper et al. (2002) also argued that segmental information, including vowel quality, is crucial for lexical activation in English. They suggested that English listeners may rely more on segmental cues than suprasegmental cues for lexical access. The argument for the role of vowel quality on word recognition is not new. A relatively earlier study by Cutler and Clifton (1984) also found that changing a reduced vowel to a full vowel or reducing a full vowel to a schwa had a negative impact on word recognition. The authors also noted that misplacement of stress suprasegmentally is also detrimental to word recognition regardless of differences in vowel quality. In the current study, there were many target words that are produced with a reduced vowel by native speakers of English (e.g., develop, consider, confusion, suspicion, position, consensus). Therefore, learners' production of these target words with a full vowel could have also led to difficulties in stress judgments carried out by the listeners. For example, saying [dɪ.vɛl.ɔp] instead of /dɪ.vɛl.əp/ may give the impression that the second and third syllables are almost equally stressed.

One other possibility could be the role of intensity in combination with vowel reduction. Although many studies ranked the role of intensity as a cue to stress lower than that of duration and pitch (e.g., Braun et al., 2011; Fry, 1955, 1958; Mattyss, 2000), intensity is still considered an important cue to primary stress. That said, both sets of acoustic analyses carried out for the current study showed that proportional intensity levels did not change over time. What is more, as Figure 7.12 shows, learners' production of syllable intensity were considerably below the native speaker baseline. Due to the limited intervention period as well as cross-linguistic transfer, learners in the current study seem to have focused on duration more than intensity and pitch in an attempt to shift their primary stress placement. However, for perceptual changes to occur, the positive gains in duration may need to be complemented with higher levels of intensity as well as correct vowel reduction for native listeners to make more confident judgments. Previous research on acoustic correlates of stress has shown that perception of stress depends on a number of acoustic cues, and if a single cue is missing, it may not affect perception negatively (Beckman, 1986; Patel et al., 2012; Turk & Sawusch, 1996). However, in the current study, there were no changes in intensity in the positive direction and no vowel reduction occurred due to L1 influence. Therefore, it may be that learners need to utilize more cues to stress then just duration to be able to reach more target-like levels in their production of primary stress. As mentioned at the beginning of this section, recasts triggered a positive change in learners' production of syllable duration. However, when the results of the acoustic analyses are considered in terms of listener judgments, it seems that the amount of acoustic gains that learners achieved did not move beyond the in-between area in the acoustic space, which made it difficult for listeners to determine the location of primary stress. Longer periods of exposure to target stress patterns through meaningful communication and more corrective feedback could possibly help learners adjust other suprasegmental and segmental aspects of their lexical stress production and move beyond the in-between zone.

That means when learners receive feedback on their production of non-target-like stress, they may attempt to change their production, but since development occurs on a continuum, the change will happen gradually. The intervention in the current study was brief but it was still able to trigger a positive shift. Since language acquisition happens over long periods of time, extended periods of exposure to correct models and more corrective feedback on errors could potentially help learners have a better control over all cues to lexical stress, which would potentially help listeners identify the location of stress easily.

Finally, despite the inconclusive results from the listener experiment, it is important to mention that there was a positive pattern in the listener data according to the manual analysis which was based on counting the number of unanimous agreements by listeners. Based on unanimous agreement by listeners, the intervention group produced only 1 word with stress on the second syllable on the pretest but 6 words on the posttest. The numbers for the control group were 2 words on the pretest and 1 word on the posttest. Although it is difficult to provide a confident interpretation based on this particular piece of information, it could possibly suggest that the changes in learners' interlanguage was most likely in the right direction.

8.2.4. Pedagogical implications

A number of pronunciation researchers have raised concerns in the past about the lack of research that focuses on L2 phonological development and how this situation has a direct impact on classroom practice as it deprives teachers from having access to research-based guidance on teaching pronunciation or dealing with pronunciation issues as they arise (Levis, 2005). Fortunately, this trend has been changing with more and more studies investigating various aspects of L2 pronunciation which have been summarized in recently published books (e.g., Derwing & Munro, 2015) and meta-analyses (e.g., Lee et al., 2015). Although this dissertation study is mainly an SLA study and was conducted in a laboratory setting without any explicit pronunciation instruction, it focuses on L2 phonological development, and as such, it has a number of pedagogical implications for language classrooms whether pronunciation is the main target of instruction or it is dealt with tangentially. In order to address concerns about ecological validity of laboratory studies conducted in the SLA field, it needs to be noted that the types of conversational interactions observed in classroom and laboratory settings are similar (Gass et al., 2005; Mackey & Goo, 2007; Russell & Spada, 2006), which makes the findings and implications of the current study relevant for classroom contexts. What is more, there are already a series of studies by Saito (2013a, 2013b, 2015a, 2015b) which showed that recasts provided in an instructional environment were able to facilitate acquisition of segmental contrasts. Therefore, whether they are provided as part of conversational interaction in dyads or in a classroom environment, there is evidence that recast can facilitate L2 phonological development. Furthermore, Sheen's (2006) classroom study showed that the type of recasts that were used in the current study are in fact very common in classroom settings. She mentioned that the majority of recasts that were provided by the teachers in the classrooms observed were "short, more likely to be declarative in mode, reduced, repeated, with a single error focus, and involve substitutions rather than deletions or additions" (pp. 386-387). The fact that the recasts provided in the current study share the same characteristics with recasts that naturally occur in classroom settings increases the ecological validity of the pedagogical implications gleaned from this study.

Studies that investigated the integration of pronunciation instruction into classroom practice have emphasized that in many cases language teachers may lack the skills to teach pronunciation (Breitkreutz, Derwing, & Rossiter, 2001; Foote, Holtby, & Derwing, 2011; MacDonald, 2002). Furthermore, teachers may struggle due to time constraints and not be able to devote a dedicated period of time to pronunciation instruction. In both of these situations, an interactionist approach to promoting pronunciation development could be a viable solution. When classroom instruction follows an interactionist approach, issues related to pronunciation can be dealt with incidentally, in other words, as they occur during communication between students or between the teacher and the students. As Long (2007) also suggests "teachers in task-based, content-based, immersion, and other kinds of second language classrooms may have the option of dealing with many of their students' language problems *incidentally* while working on their subject matter of choice" (emphasis in original, pp. 76-77). For example, if a student produces a particular word with the wrong stress pattern during a task-based activity that focuses on question formation, the teacher may simply provide a recast without interrupting the activity or changing its focus. The non-interruptive and non-intrusive nature of recasts has also been highlighted by other SLA researchers (e.g., Doughty, 2001; Gass & Mackey, 2006; Mackey, 2012). In the example given above, the focus on meaning will be maintained, the students will not deviate from the task, the teacher will not have to make extra time for providing metalinguistic explanations about primary stress, and last but not least, the student may benefit from that feedback and improve his/her pronunciation while working on grammar. The current study followed a similar approach by utilizing corrective feedback on pronunciation during a communicative task. No explicit instruction on pronunciation was provided during the intervention phase of the current study. The tasks that learners carried out during data collection were built around the goals of preparing for and conducting an interview, and at the end discussing the results of the interview. The tasks promoted brainstorming for ideas on a specific topic, exchanging information on that topic with a partner, conducting an interview to collect information about a candidate, and finally, deciding with a partner whether to hire the candidate or not. There were no instances of explicit instruction on lexical stress in English throughout the data collection period. In the same way, learners did not receive any form of metalinguistic feedback that provided even a brief definition such as how a syllable with primary stress needs to be produced with more emphasis and energy. Yet, learners were able to receive feedback on their primary stress errors while carrying out a communicative task, and based on the results of the exit survey, they noticed the corrections; and based on the results of the acoustic analysis, they utilized syllable duration to act upon the feedback they received and improve their production of primary stress. Therefore, the current study provides support for Long's (2007) point that recasts can be utilized to address learner needs that arise incidentally while focusing on other linguistic and curricular goals simultaneously. However, the findings from the current study should still be interpreted with caution as the intervention and recasts were not enough to trigger the intended shift in listener judgments of learners' lexical stress placement.

Keeping the same caveat in mind, the current study also provides implications for teachers who understand the importance of teaching pronunciation but do not have the necessary skills to do so (for a detailed discussion, see Burgess & Spencer, 2000). Teachers do not need extensive training to be able to provide recasts on learner errors. In fact, parents provide recasts on children's linguistic errors or native speakers provide recasts during conversations with non-native speakers. In these situations, recasts are a natural response by a more competent speaker, who may have no training in language teaching or linguistics, to a less competent speaker's non-target-like production. Therefore, until language teachers improve their skill set, they could also address learners' pronunciation errors with recasts and maintain recasts as part of their teaching repertoire even after they have developed a stronger skillset. Recasts would also be an ideal solution for those teachers who think overt error correction has a negative influence on learners' affective state. Some teachers may refrain from correcting students' pronunciation errors simply because they do not want students to feel embarrassed in front of their classmates (Lasagabaster & Sierra 2005;

MacDonald, 2002). However, students need corrective feedback to learn from it and improve their output to become more intelligible. Thus, in this kind of situation, recasts could possibly be utilized as they provide the correction covertly and focus on meaning rather than language.

Recasts may also be beneficial in language classes that particularly focus on pronunciation instruction. Foote et al. (2011) report that a common problem faced by teachers who teach stand-alone pronunciation classes is that it can be difficult to find aspects of pronunciation to focus on especially when learners do not share the same background, which is highly typical of ESL contexts. That said, even in EFL contexts where learners come from the same L1 background, there may be mixed-levels classrooms with learners who are at different proficiency levels and have different pronunciation needs. In both of these cases, recasts could be employed when addressing incidental pronunciation issues that are not the focus of that day's topic. For instance, during a class that focuses on voicing contrasts between the bilabial consonants /p/ and /b/, the teacher could possibly deal with a primary stress error by providing a recast.

The importance of suprasegmentals in terms of learners' intelligibility and comprehensibility has been highlighted by number of L2 pronunciation studies that focused on English as the target language (Derwing & Rossiter, 2002; Field, 2005; Hahn, 2004; Munro & Derwing, 1999). Research has also shown that lexical stress has an important role in English as well as other languages due to its role in facilitating word recognition and lexical access by the listener (Cooper et al. 2002; Cutler & van Donselaar, 2001; Jesse et al., 2017; Mattys, 2000; van Donselaar et al., 2005; Soto-Faraco et al., 2001). Therefore, it would be beneficial for learners if they received corrective feedback on their stress misplacement errors. Sometimes teachers may overlook learners' pronunciation errors thinking that they are not important enough to negatively influence the learners' intelligibility or to cause a communication breakdown (MacDonald, 2002). However, as Morley (1991) explains intelligibility is a broad term and what counts as intelligible may differ from one listener to another. It is also logical to assume that after a while teachers become accustomed to their students' pronunciation and are less likely to have comprehension problems compared to people who have limited contact with different varieties of English spoken by non-native speakers. In fact, research has shown that comprehensibility and intelligibility can be greatly affected by the listener's background as well as factors such as their attitudes toward non-native speakers (e.g., Dewaele & McCloskey, 2015; Kang & Rubin, 2009; Lindeman & Subtirelu, 2013). Therefore, not providing corrective feedback on learners' pronunciation errors, including primary stress errors, would be doing learners a disservice as they may experience communication breakdowns while interacting with speakers they meet outside the classroom. Finally, as students consider corrective feedback necessary (e.g., Huang & Jia, 2016) and early evidence from interactionist studies reveals that corrective feedback has the potential to facilitate perception of target sounds (e.g., Lee & Lyster, 2016a, 2016b; Saito & Wu; 2014) and development of segmentals (Saito & Lyster, 2012a; Saito 2013a, 2013b, 2015b) as well as suprasegmentals (Parlak & Ziegler, 2017), recasts can be an effective tool for language teachers when they are dealing with segmental and suprasegmental pronunciation errors.

8.2.5. Limitations and future directions

Although the current study showed that recasts triggered gains in duration and pitch in the direction of the native-speaker baseline, like all studies it has important limitations that will need to be redressed by future studies. The first limitation is that this study did not provide any implications regarding the amount of change that needs to occur across the main acoustic correlates of primary stress to trigger the intended perceptual judgments. As it has been mentioned earlier, there are a number of acoustic correlates of stress that join together to form the perceived prominence of a particular syllable. The current study showed that recasts had a larger positive impact on the production of duration, and to a lesser extent, on the production of pitch. However, what remains unknown is the degree of change required among the combination of acoustic cues to trigger changes in perception among various groups of listeners as well as the amount of exposure and feedback required to achieve that degree of change. Therefore, future interactionist studies that focus on the development of lexical stress would need to consider the fact that phonological development occurs on a continuum in the acoustic space and there will be a certain threshold that the combination of the acoustic cues for stress will need to pass in order for a syllable to be perceived as stressed by listeners. It will probably be difficult to determine this threshold and the amount of interaction or corrective feedback that is required to reach that threshold in a single study; therefore, a series of future studies may be necessary to understand various aspects of the relationship between these factors.

In relation to the point about phonological development occurring on a continuum, it is also important to interpret any type of development that might occur in relation to learners' proficiency level. The current study provided proficiency information in the form of IELTS and TOEFL scores; that said, it is acknowledged that learners had taken the tests a couple of weeks or months apart from one another. Despite the fact that the difference in time was limited to a couple of months, it is possible for more learning to happen even in a matter of weeks. This could be considered another limitation of the current study. In order to preempt this issue and increase the accuracy of proficiency measurements, a possible solution that could be adopted in the future is giving learners a brief proficiency test at the onset of data collection. From a logistical perspective, it would be better if this test can be taken in a short period of time. This way, there would be a second measure of proficiency to complement IELTS, TOEFL, or any other proficiency scores reported by learners.

It is also worth mentioning the inconclusive listener judgments were another limitation of the current study. Despite the possible reasons that have been discussed earlier in Chapter 8.2.3, there is no denying that agreement among listeners would have paved the way to stronger conclusions derived from the findings. In order to tackle the listener issue that was faced in the current study, further investigation with L1 speakers of English who have not had any language learning experience and have not lived in non-English speaking countries can be conducted. In fact, it would also be interesting to explore if there are any differences between listeners based on variables such as nativeness, general L2 experience, or experience in the target L2. These are some of the points that I am planning to investigate further in future studies.

One other limitation of the current study is that it does not provide insights into the changes that may have occurred in learners' perception of primary stress. Despite their differences, speech acquisition theories such as SLM (Flege, 1995, 2003) and PAM-L2 (Best & Tyler, 2007) emphasize the role of perception as a precondition to production. SLM and PAM-L2 particularly focus on the development of segmental features, and as such, they provide limited implications for research that focuses on the development of lexical stress or any other suprasegmental feature. That said, the proposition that there is a connection between perception and production would also be relevant in the case of lexical stress and other suprasegmental features. Therefore, future interactionist studies are needed to investigate whether corrective feedback leads to an improvement in L2 learners' perception of primary stress or other suprasegmental features in addition to an improvement in their production. There are already a number of interactionist studies that have explored the relationship between recasts and perception of segmental contrasts and found a positive effect (e.g., Lee & Lyster, 2016a; Saito, 2013a, 2015a). Future interactionist studies that focus on lexical stress as a target future would make valuable contributions to this line of research.

Another limitation was that there was no delayed posttest in the current study. As such, it was not possible to observe whether the immediate positive effects that were captured in the case of duration and pitch measures are durable. Mackey and Polio (2009) emphasize the importance of conducting delayed posttest and mention that "what counts as evidence of learning is open to discussion, but many researchers believe that any small changes in production, comprehension, or awareness, indicate some evidence of learning. Clearly, long-term delayed post-tests and evidence of change across a variety of contexts is the best evidence of learning" (p. 5). The current study contributed to the interactionist literature by exploring relationship between recasts and lexical stress using a task-based approach. As both the pilot study and the current study consistently showed, recasts helped L1 Arabic speakers to produce target syllables with longer duration and higher pitch based on the results of immediate posttest. However, it is important to understand whether the effects will be sustained over a longer period of time. In order to address this limitation, future research that incorporates delayed posttests into the study design will be necessary.

Although measuring the amount of noticing was not the main goal, the current study at the very least provided evidence for noticing in the form of questionnaire responses. However, noticing and attention are core constructs within the interactionist framework and they deserve more "attention" from the researcher.

According to Schmidt (1994, 2001) and Robinson (1995, 2003) noticing of new target features cannot take place without focused attention and both are necessary for language learning to occur. Therefore, providing evidence for noticing of the target features or corrective feedback allows researchers to argue that at least some form of learning took place. However, measuring noticing can be challenging as in many cases it is difficult to capture what is happening inside a learner's head during the interaction. There are a number of online and offline data collection techniques that SLA researchers employ when gauging noticing. Online techniques are used concurrently with language data collection while learners are actually carrying out the language activity or engaging in verbal or written interaction. Commonly used online techniques for measuring noticing are think-alouds (Leow, 2000) and immediate recall (Philp, 2003). Thanks to the advancements in technology, more recently eye-tracking technology has also been used as a window to the learner's mind, and hence, noticing as well as processing of language input (e.g., Godfroid, Housen, & Boers, 2010; Godfroid, Boers, & Housen, 2013; Révész et al., 2014). When it comes to offline techniques, SLA researchers generally use questionnaires (Mackey, 2006; Robinson, 1997), retrospective verbal reports (Rebuschat, 2013; Rebuschat, Hamrick, Riestenberg, Sachs, & Ziegler, 2015) and stimulated recall (Lai & Zhao, 2006; Mackey et al., 2000; Mackey, 2006). In general, offline techniques would be more feasible for measuring noticing in the case of interactionist studies that focus on L2 phonological development. As Godfroid et al. (2010) explain, online protocols "carry a risk of 'reactivity', i.e. concurrent verbalisation of cognitive processes may influence the very cognitive processes one is aiming to describe" (p. 174). Online protocols make things more complicated when the target language feature is a phonological feature and the goal is to maintain a natural conversation during data collection. Therefore, future studies could make use of stimulated recall in addition to questionnaires when measuring noticing. Needless to say, it is highly crucial to present the stimulus (e.g., video recording) immediately after the interaction session so that learners are able to access the short-term memory when explaining what they see in the stimulus (Mackey & Gass, 2016). Triangulation of introspective methods would yield data that gives better insights into what aspects of the phonological feedback learners actually attend to, which is among the limitations of the current study. As it has been mentioned earlier, there is already evidence that learners understand the intent of corrective feedback when it is directed toward pronunciation issues. That said, stimulated recall could help us understand whether learners correctly interpret the target of the feedback (e.g., lexical stress) in addition to noticing that it is corrective feedback on pronunciation. The questionnaire used in the current study showed that learners noticed that their pronunciation was being corrected, but since no learner explicitly mentioned stress, whether they noticed the specific target of the feedback remains unknown.

The current study did not provide any insights on the effects of recasts on untrained items, which is another limitation. Saito (in press) emphasizes the relationship between vocabulary learning and sound learning, and argues that researchers should not only measure phonetic development by focusing on the trained lexical items but also measure whether learners are able to generalize their gains in phonetic realization to new vocabulary. Saito's recommendation arises within the context of research studies that are designed to provide explicit instruction, particularly form-focused instruction, coupled with corrective feedback. The current study was a laboratory study that did not have an instructional component. That said, learners' ability to generalize their new knowledge to untrained lexical items is still relevant in the case of laboratory studies. The current study focused on a particular set of target words when providing recasts as well as when measuring the development of primary stress placement. However, it did not investigate the changes that may have occurred in untrained vocabulary. Therefore, further research is necessary to understand whether corrective feedback leads to a positive shift in the acoustic realization of lexical stress for untrained words in addition to trained words. It is important to highlight that newly-acquired phonological or phonetic knowledge will be generalized differently depending on whether the target feature is a segmental one or a suprasegmental one. For example, when a learner masters the production of phonemic contrasts between problematic pairs such as /p/-/b/ in the case of L1 Arabic speakers or /ɹ/-/l/ in the case of Japanese speakers, it is easy to test whether they can make use of their newly-acquired knowledge when they produce other words as English is abound with words that have these phonemes in various positions. On the other hand, measuring generalizations based on knowledge of lexical stress is not a straight-forward process. Apart from some exceptions, the location of primary stress is mainly unpredictable in English and it needs to be learned on a case-by-case basis similar to learning word meaning. Therefore, learning the stress pattern of one particular word does not necessarily mean that learners will be able to apply this knowledge to correctly produce the stress patterns of a new word at first encounter. That said, what could be measured using untrained words is whether learners are able to generalize the notion of primary stress. This would particularly be relevant in the case of learners whose L1 is a syllable-timed language. As a first step, research can determine whether corrective feedback facilitates the development of the notion of stress as a phonological component in the learner's interlanguage, which can be measured through phonetic analysis of a group of target words. After that, learners can be tested on untrained words in order to understand whether they are able to generalize their newly-acquired phonological knowledge, which is the notion that one syllable in any multisyllabic word needs be produced with more emphasis compared to its neighboring syllables, and attempt to produce primary stress when they encounter a new word. Even if their primary stress placement is on the wrong syllable, the attempt to produce some form of primary stress through the manipulation of acoustic cues would indicate that they have understood the concept of primary stress. In a similar vein, future studies could investigate generalization of phonological knowledge such as vowel reduction when L1 speakers of a stress-timed language that does not have vowel reduction, such as Arabic, acquire lexical stress in a new language, such as English.

There are also a few important implications that are gained from the current study. One of them is related to the length of treatment sessions. According to Long (2007), explicit types of feedback such as metalinguistic feedback or explicit correction work better when the treatment is short-term and implicit feedback types such as recasts work well when there are opportunities for longer exposure. The current study provided opportunities for recasts in a relatively brief period of time. Despite the brief intervention period, the results showed that recasts led to gains in syllable duration consistent across intervention group learners and the words that they produced. Therefore, a possible interpretation could be that the efficiency of recasts depends on the target language feature in addition to the amount of exposure learners receive. This is not surprising considering previous research findings which showed that learners are able to notice recasts that target phonological errors more than recasts that target morphological errors (Carpenter et al., 2006; Mackey et al., 2000) and that gains are directly related to quality and intensity of input (Norris & Ortega, 2000).

That is to say, the positive effects of recasts may be observed in a shorter period of time when the target of feedback is a phonological feature.

Another implication is the need for triangulation of the outcome measures in studies that focus on L2 phonological development. As the current study showed, relying on listener judgments only would have given the impression that learners had not benefitted from corrective feedback at all. On the other hand, relying only on acoustic analyses would have led to the interpretation that there was a positive change in learners' production of lexical stress but it would not have revealed that it was not enough to be picked up by the listeners. Using both acoustic analysis and listener judgments painted a clearer picture and led to more careful interpretation of the results. Although the current study is not the first SLA study that uses both acoustic analyses and listener judgments as the outcome measures, pronunciation and SLA literature are abound with studies that use only acoustic analyses or only listener judgments for the analysis of data. The results of the current study suggest that for future studies intending to focus on L2 phonological development, intelligibility, and/or comprehensibility, researchers may want to consider using both acoustic analyses and listener judgments for triangulation purposes. Carrying out acoustic analyses can be very time consuming depending on the size of the data; however, the extra efforts would be justified by increased confidence in the results as well as the interpretation that follows. The call for triangulation and increased role of acoustic analyses has in fact been made by other scholars who underscored the importance of employing acoustic analyses in applied linguistics studies (e.g., Kang, Rubin, & Pickering, 2010; Kang & Ginther, 2018).

The current study offers future directions for research endeavors that would help us understand more about the impact of interaction and corrective feedback on the development of lexical stress. L1 Arabic speakers in this study used duration in an attempt to improve their primary stress placement. As duration is the major cue for stress in Arabic, this finding provides evidence for cross-linguistic influence. However, it would also be interesting to study the developmental patterns of lexical stress by L1 speakers of syllable-timed languages such as Turkish and Japanese. This would help understand the impact of interaction and corrective feedback on the development of lexical stress when there is no possibility for cross-linguistic transfer. Do L1 speakers of non-stress languages also rely on duration because it is the most commonly utilized correlate? Or do they utilize other correlates such as intensity or pitch? Are there any differences in the developmental patterns of lexical stress depending on which syllable-timed first language learners speak? For example, do Turkish and Japanese learners react differently to recasts on lexical stress errors? These are questions that future studies can investigate.

Another area of investigation could be the impact of corrective feedback when it is provided as part of pronunciation instruction. Recent research on the acquisition of segmental contrasts has shown that the combination of form-focused instruction and recasts is beneficial for learners (e.g., Saito & Lyster, 2012a, 2013b; Saito, 2013a, 2013b, 2015a, 2015b). However, interactionist studies focusing on L2 phonological development are very few in number, and to the best of my knowledge, there is no interactionist study that has investigated the impact of form-focused instruction on the development of lexical stress. Therefore, this is an area that waits to be addressed.

Finally, future studies that focus on the development of lexical stress could investigate the differential effects of cognitive factors on learners' success rates. Previous research has suggested that cognitive factors such as phonemic coding ability, phonological memory and working memory can have a positive impact on L2 development. SLA research to date has investigated various types of relationships such as working memory capacity and noticing of interactional feedback (e.g., Mackey, Philp, Egi, Fujii & Tatsumi, 2002; Trofimovich et al., 2007), working memory capacity and interactional feedback type (Yilmaz, 2013), working memory and effectiveness of feedback (Révész, 2012), phonemic coding and phonological accuracy, rote memory and fluency (Saito, 2017), and phonological memory and fluency (O'Brien, Segalowitz, Freed, & Collentine, 2007). Although more research is necessary to be able to make confident associations, there seems to be a positive relationship between working memory and noticing as well as phonological memory and fluency. Also, a recent study by Li and DeKeyser (2017) found a positive correlation between musical tonal ability and perception and production of Mandarin tones by L2 learners. Many of these cognitive factors are relevant in the case of lexical stress. Therefore, it would be interesting to see future studies that investigate the effects of phonological memory, working memory, and musical ability on the development of lexical stress.

Overall, the main goal of this dissertation was to explore an understudied area and expand our understanding of how interaction and corrective feedback can facilitate development of a suprasegmental feature. Needless to say, this study has raised many more questions compared to the ones that it has answered. Hopefully, findings and implications of this dissertation will spark an interest among other SLA scholars to pursue some of these questions. Investigating the development of suprasegmental features from an interactionist perspective is an interesting and a relatively new area. It has the potential to make meaningful contributions to the debate surrounding the effects of interaction and corrective feedback on language development. Due to these reasons, it is where my future research agenda will focus.

In a series of follow-up studies, I am planning to investigate the development of lexical stress by Thai and Turkish speakers. These future studies will provide an opportunity to answer new questions related to the development of lexical stress as well as to address some of the limitations mentioned in this section. As my research agenda continues to develop in the upcoming years, it will most likely expand to include the study of other suprasegmental features. It is my hope that the current dissertation and the follow-up studies I am planning to conduct would motivate future empirical investigations and deepen our understanding of how languages are acquired.

REFERENCES

Abercrombie, D. (1949). Teaching pronunciation. *English Language Teaching, 3*, 113-122.

Abercrombie, D. (1991). *Fifty years in phonetics: Selected papers*. Edinburgh: Edinburgh University Press.

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research, 7*(3), 347-376.

Akita, M. (2005). The effectiveness of a prosody-oriented approach in L2 perception and pronunciation training. *Academic Studies, English Language and Literature, 53*, 1-22.

Almbark, R., Bouchhioua, N., & Hellmuth, S. (2014). Acquiring the phonetics and phonology of English word stress: Comparing learners from different L1 backgrounds. *Concordia Working Papers in Applied Linguistics*, *5*, 19-35.

Ammar, A., & Spada, N. (2006). One size fits all? Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition, 28*(4), 543-574.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42(4),* 529-555.

Archibald, J. (1992). Transfer of L1 parameter settings: Some evidence from Polish metrics. *Canadian Journal of Linguistics, 37*, 301-339.

Archibald, J. (1993). The learnability of English metrical parameters by adult Spanish speakers. *International Review of Applied Linguistics, 31,* 129-141.

Archibald, J. (1994). A formal model of learning L2 prosodic phonology. *Second Language Research*, *10*(3), 215-240.

Archibald, J. (1997). The acquisition of English stress by speakers of nonaccentual languages: Lexical storage versus computation. *Linguistics, 35(1)*, 167-181.

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24*, 621-63.

Arvaniti, A. (2000). The phonetics of stress in Greek. *Journal of Greek Linguistics*, *1*(1), 9-39.

Arvaniti, A. (in press). The autosegmental metrical model of intonation phonology. In S. Shattuck-Hufnagel, & J. Barnes (Eds.), *Prosodic theory in practice*. Cambridge, MA: MIT Press.

Atal, B. S. (1968). *Automatic speaker recognition based on pitch contours* (Doctoral dissertation). Polytechnic Institute of Brooklyn, NY.

Audacity Team (2018). Audacity(R): Free audio editor and recorder (Version 2.1.0) [Computer software]. Retrieved from https://audacityteam.org/

Ayoun, D. (2001). The role of negative and positive feedback in the second language acquisition of the passé composé and imparfait. *The Modern Language Journal*, *85*(2), 226-243.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Baker, W., & Trofimovich, P. (2006). Perceptual paths to accurate production of L2 vowels: The role of individual differences. *International Review of Applied Linguistics in Language Teaching, 44*, 231-250.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition, 35*(4), 689-725.

Barkat, M., Ohala, J., & Pellegrino, F. (1999). Prosody as a distinctive feature for the discrimination of Arabic dialects. *Sixth European Conference on Speech Communication and Technology,* 395-398.

Barr, D. J. (2013). Random effects structures for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4,* 1-2.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255-278.

Basetti, B. (2006). Orthographic input and phonological representations in learners of Chinese as a foreign language. *Written Language and Literacy, 9*(1), 95-114.

Basetti, B., & Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Applied Psycholinguistics, 36,* 67-91.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.*

Bates, E., & MacWhinney, B. (1981). Second-language acquisition from a functionalist perspective: Pragmatic, semantic, and perceptual strategies. *Annals of the New York Academy of Sciences*, *379*(1), 190-214.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Beckman, M. (1986). *Stress and non-stress accent.* Dordrecht: Foris Publications.

Best, C. T. (1995). A direct-realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 171-204). Timonium, MD: York Press.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology:*

Human Perception and Performance, 14(3), 345-360.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception. In O. S. Bohn, & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-24). Amsterdam: John Benjamins.

Bion, R. A. H., Escudero, P., Rauber, A. S., & Baptista, B. O. (2006). Category formation and the role of spectral quality in the perception and production of English front vowels. *Proceedings of Interspeech 2006, 1363-1366.*

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning, 56*, 9-49.

Bitchener, J. (2004). The relationship between the negotiation of meaning and language learning: A longitudinal study. *Language Awareness*, *13*(2), 81-95.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences, 17*, 97-119.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer (Version 6.0.28) [Computer program]. Retrieved from http://www.praat.org/

Bohn, O. S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, *11*(3), 303-328.

Bohn, O. S., & Flege, J. E. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition*, *14*(2), 131-158.

Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, *14*, 109-149.

Bond, Z. S. (1999). Morphological errors in casual conversation. *Brain and Language,* *68*, 144-150.

Bond, Z. S. (2005). Slips of the ear. In D. Pisoni, & R. Remez (Eds.), *The handbook of speech perception* (pp. 290-310). Malden, MA: Blackwell Publishing.

Bond, Z. S., & Small, L. H. (1983). Voicing, vowel, and stress mispronunciations in continuous speech. *Perception & Psychophysics*, *34*(5), 470-474.

Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /ɹ /–/l/ contrast. In J. G. H. Edwards, & M. L. Zampini (Eds.), *Phonology in Second Language Acquisition*, (pp. 287-308). Amsterdam: John Benjamins.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707-729.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310.

Braidi, S. M. (2002). Reexamining the role of recasts in native-speaker/nonnative-speaker interactions. *Language Learning*, *52*(1), 1-42.

Braine, M. D. S. (1971). On two types of models of the internalization of grammars. In D. I. Slobin (Ed.), *The ontogenesis of grammar: A theoretical symposium* (pp. 153–186). New York, NY: Academic Press.

Braun, B., Lemhöfer, K., & Mani, N. (2011). Perceiving unstressed vowels in foreign-accented English. *The Journal of the Acoustical Society of America*, *129*(1), 376-387.

Breen, M. P. (1987). Learner contributions to task design. In C. N. Candlin, & D. Murphy (Eds.), *Lancaster practical papers in English language education:*

Vol. 7. language learning tasks (pp. 23-46). Englewood Cliffs, NJ: Prentice Hall.

Breen, M. P. (1989). The evaluation cycle for language learning tasks. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 187-206). New York, NY: Cambridge University Press.

Breitkreutz, J., Derwing, T. M., & Rossiter, M. J. (2001). Pronunciation teaching practices in Canada. *TESL Canada Journal, 19*(1), 51-61.

Brock, C., Crookes, G., Day, R., & Long, M. (1986). Differential effects of corrective feedback in native speaker-nonnative speaker conversation. In R. R. Day (Ed.), *Talking to learn*: *Conversation in second language acquisition* (pp. 229-236). Rowley, MA: Newbury House.

Broselow, E., Chen, S., & Wang, C. (1998). The emergence of the unmarked in second language phonology. *Studies in Second Language Acquisition, 20,* 261-280.

Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics, 34,* 419-441.

Brown, D. (2016). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research*, *20*(4), 436-458.

Browning, S. R. (2004). *Analysis of Italian children's English pronunciation.* Report contributed to the EU FP5 PF STAR Project. Retrieved from http://archive.is/zsxA

Bueno-Alastuey, M. C. (2013). Interactional feedback in synchronous voice-based computer mediated communication: Effect of dyad. *System, 41*, 543-559.

Burgess, J., & Spencer, S. (2000). Phonology and pronunciation in integrated language teaching and teacher education. *System*, *28*(2), 191-215.

Busà, M. G., & Urbani, M. (2011). A cross linguistic analysis of pitch range in English L1 and L2. *Proceedings of the 17th International Congress of Phonetic Sciences,* 380-383.

Bygate, M., Skehan, P. & Swain, M. (2001). Introduction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 1-20). Harlow: Pearson Education.

Byrnes, J. (2008, April 1). Re: Coefficient of determination (R^2) when using lme() [Electronic mailing list message]. Retrieved from https://stat.ethz.ch/pipermail/r-sig-mixed-models/2008q2/000713.html

Camacho, A. (2007). *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music* (Doctoral dissertation). University of Florida, FL.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201-219.

Carpenter, H., Jeon, K. S., MacGregor, D., & Mackey, A. (2006). Learners' interpretations of recasts. *Studies in Second Language Acquisition, 28,* 209-236.

Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback. *Studies in Second Language Acquisition*, *15*(3), 357-386.

Cenoz, J., & Lecumberri, L. G. (1999). The acquisition of English pronunciation: Learners' views. *International Journal of Applied Linguistics, 9,* 3-17.

Chen, Y., Robb, M. P., Gilbert, H. R., & Lerman, J. W. (2001). A study of sentence stress production in Mandarin speakers of American English. *The Journal of the Acoustical Society of America*, *109*(4), 1681-1690.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637-669.

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1468-1479.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335-359.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647-3658.

Colantoni, L., Marasco, O., Steele, J., & Sunara, S. (2014). Learning to realize prosodic prominence in L2 French and Spanish. In R. T Miller et al. (Eds.), *Selected proceedings of the 2012 second language research forum: Building bridges between disciplines* (pp. 15-29). Somerville, MA: Cascadilla Proceedings Project.

Colantoni, L., Steele, J., & Escudero, P. (2015). *Second language speech: Theory and practice.* Cambridge: Cambridge University Press.

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching*, *5*(4), 161-170.

Corder, S. P. (1982). *Error analysis and interlanguage*. Oxford: Oxford University Press.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech*, *45*(3), 207-228.

Couper, G. (2003). The value of an explicit pronunciation syllabus in ESOL teaching.

Prospect, 18(3), 53-70.

Couper-Kuhlen, E. (1996). The prosody of repetition: On quoting and mimicry. In E. Couper-Kuhlen, & M. Selting (Eds.), *Prosody in conversation: Interactional studies,* (pp. 366-405). Cambridge: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369-382.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159-181). New York, NY: Routledge.

Cutler, A. (1984). Stress and accent in language production and understanding. D. Gibbon, & H. Richter (Eds.), *Intonation, accent, and rhythm: Studies in discourse phonology* (pp. 77-90). Berlin: Walter de Gruyter.

Cutler, A. (2005). Lexical stress. In D. B. Pisoni, & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 264-289). Oxford: Blackwell Publishing.

Cutler, A., & Clifton Jr, C. (1984). The use of prosodic information in word recognition. In H. Bouma, & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 183-196). Hillsdale, NJ: Lawrence Erlbaum.

Cutler, A., & van Donselaar, W. (2001). Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech*, *44*(2), 171-195.

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/

de Jong, K. (2004). Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics*, *32*(4), 493-516.

de Jong, K., & Zawaydeh, B. A. (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics*, *27*(1), 3-22.

de Jong, K., & Zawaydeh, B. A. (2002). Comparing stress, lexical focus, and segmental focus: Patterns of variation in Arabic vowel duration. *Journal of Phonetics, 30*, 53-75.

de la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition, 24*(1), 81-112.

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125-152). Cambridge: Cambridge University Press.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 20*, 1-16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*(3), 379-397.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*, 476-490.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163-185.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* Amsterdam: John Benjamins.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393-410.

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System, 30*, 155-166.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning, 13*(1), 1-17.

Deterding, D. (2001). The measurement of rhythm: A comparison of Singapore and British English. *Journal of Phonetics, 29.* doi:10.1006/jpho.2001.0138

Dewaele, J. M., & McCloskey, J. (2015). Attitudes towards foreign accents among adult multilingual language users. *Journal of Multilingual and Multicultural Development, 36*(3), 221-238.

Dogil, G., & Williams, B. (1999). The phonetic manifestation of word stress. In H. V. D. Hulst (Ed.), *Word prosodic systems in the languages of Europe* (pp. 273-334). New York, NY: Mouton de Gruyter.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-257). Cambridge, Cambridge University Press.

Doughty, C., & Valera, E. (1998). Communicative focus on form. In C. Doughty, & J. Williams (Eds.), *Focus on form in second language acquisition* (pp. 114-138). Cambridge: Cambridge University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.

Duanmu, S. (2008). *Syllable structure: The limits of variation.* Oxford: Oxford University Press.

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two*. New York: Oxford University Press.

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress “deafness”. *The Journal of the Acoustical Society of America*, *110*(3), 1606-1618.

Eckman, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning*, *27*(2), 315-330.

Eckman, F. R. (1981). On predicting phonological difficulty in second language acquisition. *Studies in Second Language Acquisition*, *4*(1), 18-30.

Eckman, F. R. (1991). The structural conformity hypothesis and the acquisition of consonant clusters in the interlanguage of ESL learners. *Studies in Second Language Acquisition*, *13*(1), 23-41.

Eckman, F. R., & Iverson, G. K. (2013). The role of native language phonology in the production of L2 contrasts. *Studies in Second Language Acquisition*, *35*(1), 67-92.

Educational Testing Service. (2010). Linking TOEFL iBT scores to IELTS scores: A research report. Retrieved from https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf

Egi, T. (2007). Recasts, learners' interpretations, and L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 249-267). Oxford: Oxford University Press.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164-194.

Ellis, N. C. (2016). Salience, cognition, language complexity, and complex adaptive systems. *Studies in Second Language Acquisition*, *38*(2), 341-351.

Ellis, N. C. (2018). Salience in usage-based SLA. In S. M. Gass, S., P. Spinner, & J. Behney, *Salience in second language acquisition* (pp. 21-40). London: Taylor and Francis.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

Ellis, R. (2007). The differential effects of corrective feedback on two grammar structures. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 339-360). Oxford: Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition.* Oxford: Oxford University Press.

Ellis, R. (2017). Oral corrective feedback in L2 classrooms: What we know so far. In H. Nassaji, & E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning: Research, theory, applications, implications* (pp. 3-18). New York, NY: Routledge.

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning*, *51*(2), 281-318.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition, 28*(02), 339-368.

Ellis, R., & Sheen, Y. (2006). Reexamining the role of recasts in second language acquisition. *Studies in Second Language Acquisition*, *28*(4), 575-600.

Embarki, M. (2013). Phonetics. In J. Owens (Ed.), *The Oxford handbook of Arabic linguistics* (pp. 23-44). New York, NY: Oxford University Press.

Erlam, R., & Loewen, S. (2010). Implicit and explicit recasts in L2 oral French interaction. *Canadian Modern Language Review*, *66*(6), 877-905.

Farrar, M. J. (1990). Discourse and the acquisition of grammatical morphemes. *Journal of Child Language*, *17*, 607-624.

Farrar, M. J. (1992). Negative evidence and grammatical morpheme acquisition. *Developmental Psychology*, *28*(1), 90-98.

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, *8*(3), 505-520.

Ferguson, C. A. (1981). 'Foreigner talk' as the name of a simplified register. *International Journal of the Sociology of Language*, *1981*(28), 9-18.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399-423.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*(1), 47-65.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America, 93*(3), 1589-1608.

Flege, J. E. (1995). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*(4), 425-442.

Flege, J. E. (2003) Assessing constraints on second-language segmental production and perception. In A. Meyer, & N. Schiller (Eds.), *Phonetics and phonology in language comprehension and production, differences and similarities* (pp. 319-355). Berlin: Mouton de Gruyter.

Flege, J. E., & Bohn, O. S. (1989). An instrumental study of vowel reduction and stress placement in Spanish-accented English. *Studies in Second Language*

Acquisition, *11*(1), 35-62.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437-470.

Flege, J. E., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech*, *24*(2), 125-146.

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2011). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada Journal*, *29*(1), 1-22.

Foote, J. A., & Trofimovich, P. (2018). Second language pronunciation learning: An overview of theoretical perspectives. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 75-90). London: Routledge.

Fox, A. (2000). *Prosodic features and prosodic structure: The phonology of suprasegmentals*. Oxford: Oxford University Press.

Fraser, S., Gagné, J. P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*, *53*, 18-33.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, *27*(4), 765-768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*(2), 126-152.

Fudge, E. (2016). *English word-stress*. New York, NY: Routledge.

Gallego, J. C. (1990). The intelligibility of three nonnative English-speaking teaching assistants: An analysis of student-reported communication breakdowns. *Issues in Applied Linguistics*, *1*(2), 219-237.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). Irr: Various coefficients of interrater reliability and agreement. R package version 0.84.

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, *1*(2), 292-304.

Gass, S. M. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, *9*(2), 198-217.

Gass, S. M. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S. M. (2003). Input and interaction. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 224-255). Malden, MA: Blackwell Publishing.

Gass, S. M. (2018). *Input, interaction, and the second language learner.* New York, NY: Routledge.

Gass, S. M., & Alvarez-Torres, M. J. (2005). Attention when? An investigation of the ordering effect of input and interaction. *Studies in Second Language Acquisition*, *27*(1), 1-31.

Gass, S. M., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA Review*, *19*(1), 3-17.

Gass, S. M., & Mackey, A. (2015). Input, interaction, and output in second language acquisition. In B. VanPatten, & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 180-206). New York, NY: Routledge.

Gass, S. M., Mackey, A., & Pica, T. (1998). The role of input and interaction in second language acquisition: Introduction to the special issue. *The Modern Language Journal*, *82*(3), 299-307.

Gass, S. M., Mackey, A., & Ross-Feldman, L. (2005). Task-based interactions in classroom and laboratory settings. *Language Learning*, *55*(4), 575-611.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65-87.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/ hierarchical models*. Cambridge: Cambridge University Press.

Gerhard, D. (2003). *Pitch extraction and fundamental frequency: History and current techniques*. Regina: University of Regina.

German, J., Pierrehumbert, J. B., & Kaufmann, S. (2006). Evidence for phonological constraints on nuclear accent placement. *Language*, *82*(1), 151-168.

Ghazali, S., Hamdi, R., & Barkat, M. (2002). Speech rhythm variation in Arabic dialects. *Proceedings of the Speech Prosody 2002 Conference*, 127-132.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*(3), 483-517.

Godfroid, A., Housen, A., & Boers, F. (2010). A procedure for testing the noticing hypothesis in the context of vocabulary acquisition. In M. Pütz, & L. Sicola (Eds.), *Cognitive processing in second language acquisition* (pp. 169–197). Amsterdam: John Benjamins.

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition, 34*, 445-474.

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 443-482). Philadelphia, PA: John Benjamins.

Goo, J., & Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition*, *35*(1), 127-165.

Gooch, R., Saito, K., & Lyster, R. (2016). Effects of recasts and prompts on L2 pronunciation development: Teaching English /ɹ/ to Korean adult EFL learners. *System, 60*, 117-127

Gordon, J., & Darcy, I. (2016). The development of comprehensible speech in L2 learners: A classroom study on the effects of short-term pronunciation instruction. *Journal of Second Language Pronunciation, 2*(1), 56-92.

Gordon, M. (2014). Disentangling stress and pitch-accent: A typology of prominence at different prosodic levels. In H. van der Hulst (Ed.), *Word stress: Theoretical and typological issues* (pp. 83-118). Cambridge: Cambridge University Press.

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1).

Gordon-Salant, S., Yeni-Komshian, G. H., Pickett, E. J., & Fitzgibbons, P. J. (2016). Perception of contrastive bi-syllabic lexical stress in unaccented and accented words by younger and older listeners. *The Journal of the Acoustical Society of America*, *139*(3), 1132-1148.

Greenberg, J. H. (1966). Synchronic and diachronic universals in phonology. *Language*, *42*(2), 508-517.

Gregg, K. R. (1996). The logical and developmental problems of second language acquisition. In W. Ritchie, & T. Bhatia (Eds.), *Handbook of second language* *acquisition* (pp. 49-81). San Diego, CA: Academic Press.

Guion, S. G., Flege, J. E., Liu, S. H., & Yeni-Komshian, G. H. (2000). Age of learning effects on the duration of sentences produced in a second language. *Applied Psycholinguistics*, *21*(2), 205-228.

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and late Spanish-English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition*, *7*(3), 207-226.

Gut, U. (2009). *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German.* Frankfurt: Peter Lang.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201–223.

Hammerly, H. (1982). Contrastive phonology and error analysis. *International Review of Applied Linguistics in Language Teaching*, *20*, 17-32.

Han, Z. (2002). A study of the impact of recasts on tense consistency in L2 output. *TESOL Quarterly*, 543-572.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *8*(1), 34-52.

Hatch, E. M. (1978). *Second language acquisition: A book of readings*. Rowley, MA: Newbury House.

Heinz, J. (2014). Culminativity times harmony equals unbounded stress. In H. van der Hulst (Ed.), *Word stress: Theoretical and typological issues* (pp. 255-275). Cambridge: Cambridge University Press.

Hellmuth, S. (2013). Phonology. In J. Owens (Ed.), *The Oxford handbook of Arabic linguistics* (pp. 45-70). New York, NY: Oxford University Press.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, *45*(3), 573-584.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346-363.

Huang, X., & Jia, X. O. (2016). Corrective feedback on pronunciation: Students' and teachers' perceptions. *International Journal of English Linguistics*, *6*(6), 245-254.

Hulstijn, J. H., & Marchena, E. (1989). Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition*, *11*(3), 241-255.

Hyman, L. M. (1977). On the nature of linguistic stress. In L. Hyman (Ed.), *Studies in stress and accent, Southern California occasional papers in linguistics 4.* Los Angeles, CA: University of Southern California.

Hyman, L. M. (2006). Word-prosodic typology. *Phonology*, *23*(2), 225-257.

Hyman, L. M. (2009). How (not) to do phonological typology: The case of pitch-accent. *Language Sciences*, *31*(2-3), 213-238.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.) *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Linguistics*, *37*, 1175-1199.

Isaacs, T. (2009). Integrating form and meaning in L2 pronunciation instruction. *TESL Canada Journal*, *27*(1), 1-12.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, *34*(3), 475-505.

Ishida, M. (2004). Effects of recasts on the acquisition of the aspectual form -te i-(ru) by learners of Japanese as a foreign language. *Language Learning, 54*(2), 311-394.

Iwashita, N. (2003). Negative feedback and positive evidence in task-based interaction: Differential effects on L2 development. *Studies in Second Language Acquisition, 25*(1), 1-36.

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in Second Language Acquisition*, *24*(4), 541-577.

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *TESOL Quarterly, 34*(2), 239-278.

James, A. R. (1986). *Suprasegmental phonology and segmental form.* Germany: De Gruyter.

Jassem, W., & Gibbon, D. (1980). Re-defining English accent and stress. *Journal of the International Phonetic Association*, *10*, 2-16.

Jenkins, J. (2004). Research in teaching pronunciation and intonation. *Annual Review of Applied Linguistics*, *24*, 109-125.

Jesse, A., Poellmann, K., & Kong, Y. Y. (2017). English listeners use suprasegmental cues to lexical stress early during spoken-word recognition. *Journal of Speech, Language, and Hearing Research*, *60*(1), 190-198.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944-946.

Jun, S. -A. (2005). Prosodic typology. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 439-458). Oxford: Oxford University Press.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*(2), 301-315.

Kang, O., & Ginther, A. (2018). Introduction. In O. Kang, & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 14-22). New York, NY: Routledge.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28*(4), 441-456.

Kang, O., Rubin, D. L., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554-566.

Katamba, F. (1996). *An introduction to phonology.* London: Longman.

Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Philadelphia, PA: Benjamins.

Keelan, E., Lai, C., & Zechner, K. (2010, November). The importance of optimal parameter setting for pitch extraction. *Proceedings of Meetings on Acoustics, Acoustical Society of America, 11*. doi: 10.1121/1.3609833

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 Speech: The role of listener experience and semantic context. *The Canadian Modern Language Review, 64*(3), 459-489.

Kim, J., & Han, Z. (2007). Recasts in communicative EFL classes: Do teacher intent and learner interpretation overlap? In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 269-297). Oxford: Oxford University Press.

Kim, Y. (2017). Cognitive-interactionist approaches to L2 instruction. In S. Loewen & S. Masatoshi (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 126-145). New York, NY: Routledge.

Kirkham, S. (2014). Praat script for extracting marked words.

Kirkham, S. (2015). Praat script for extracting acoustic measures from words.

Kirkham, S. (2017). Ethnicity and phonetic variation in Sheffield English liquids. *Journal of the International Phonetic Association, 47*(1), 17-35.

Kissling, E. M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal, 97*(3), 720-744.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, *118*(2), 1038-1054.

Krashen, S. D. (1977). Some issues relating to the monitor model. In H. Brown, C. Yorio, & R. Crymes (Eds.), *On TESOL '77* (pp. 144-158). Washington, D.C.: TESOL.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon Press.

Krashen, S. D. (1983). Newmark's "ignorance hypothesis" and current second language acquisition theory. In S. M. Gass, & L. Selinker (Eds.), *Language transfer in language learning* (pp. 135-153). New York, NY: Newbury House.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. London: Longman.

Krashen, S. D. (1994). The input hypothesis and its rivals. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 45-77). London: Academic Press.

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge: Cambridge

University Press.

Ladefoged, P. (2006). *A course in phonetics* (5th ed.). Boston, MA: Thomson Wadsworth.

Ladefoged, P., & Disner, S. F. (2012). *Vowels and consonants* (3rd ed.). Oxford: John Wiley & Sons.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers.* Ann Arbor, MI: The University of Michigan Press.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*(4), 433-451.

Lai, C., & Zhao, Y. (2006). Noticing and text-based chat. *Language Learning & Technology, 10*(3), 102-120.

Lai, Y., Sereno, J., & Jongman, A. (2008). Acoustic realization and perception of English lexical stress by Mandarin learners. *Canadian Acoustics, 36*(3), 124-125.

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics, 26*, 227-247.

Lantolf, J. (2000). *Sociocultural theory and second language learning*. Oxford: Oxford University Press.

Larsen-Freeman, D. (2007). Reflecting on the cognitive-social debate in second language acquisition. *The Modern Language Journal, 91*, 773-787.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R.* New York, NY: Routledge.

Lasagabaster, D., & Sierra, J. M. (2005). Error correction: Students' versus teachers' perceptions. *Language Awareness*, *14*(2-3), 112-127.

Leather, J. (1983). Second-language pronunciation learning and teaching. *Language Teaching*, *16*(3), 198-219.

Lee, A. H., & Lyster, R. (2016a). The effects of corrective feedback on instructed L2 speech perception. *Studies in Second Language Acquisition*, *38*(1), 35-64.

Lee, A. H., & Lyster, R. (2016b). Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning*, *66*(4), 809-833.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345-366.

Leeman, J. (2003). Recasts and second language development. *Studies in Second Language Acquisition*, *25*(1), 37-63.

Leeman, J. (2007). Feedback in L2 learning: Responding to errors during practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from linguistics and psychology* (pp. 111-137). Cambridge: Cambridge University Press.

Leemann, A. (2012). *Swiss German intonation patterns*. Amsterdam: John Benjamins. Philadelphia: Benjamins, to appear 2012 Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*(5), 573-579.

Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *The Journal of the Acoustical Society of America*, *31*(4), 428-435.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software, 69*(1), 1-33.

Lenth, R. V. (2018) *Using lsmeans*. Retrieved from https://cran.r-project.org/web/packages/lsmeans/vignettes/using-lsmeans.pdf

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition, 22*(4), 557-584.

Leow, R. P. (2002). Models, attention, and awareness in SLA: A response to Simard and Wong's "Alertness, orientation, and detection: The conceptualization of attentional functions in SLA." *Studies in Second Language Acquisition*, *24*(1), 113-119.

Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York, NY: Routledge.

Leow, R. P. & Martin, A. (2018). Enhancing the input to promote salience of the L2. In S. M. Gass, S., P. Spinner, & J. Behney, *Salience in second language acquisition* (pp. 167-186). London: Taylor and Francis.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369-377.

Levis, J. M., & Sonsaat, S. (2018). Pronunciation teaching in early CLT era. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 267-283). London: Routledge.

Levis, J. M., & Wu, A. (2018). Pronunciation: Research into practice and practice into research. *CATESOL Journal*, *30*, 1-12.

Lewis, C., & Deterding, D. (2018). Word stress and pronunciation teaching in English as a lingua franca contexts. *CATESOL Journal, 30*(1), 161-176.

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition, 39*(4), 593-620.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*(2), 309-365.

Li, S. (2014). The interface between feedback type, L2 proficiency, and the nature of the linguistic target. *Language Teaching Research*, *18*(3), 373-396.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8*(2), 249-336.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America*, *32*(4), 451-454.

Lightbown, P. M., & Spada, N. (1990). Focus-on-form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition*, *12*(4), 429-448.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185-207.

Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning*, *63*(3), 567-594.

Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, *39*(4), 723-730.

Loewen, S. (2009). Recasts in multiple response focus on form episodes. In A. Mackey, & C. Polio (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 176-196). New York, NY:

Routledge.

Loewen, S., & Erlam, R. (2006). Corrective feedback in the chatroom: An experimental study. *Computer Assisted Language Learning*, *19*(1), 1-14.

Loewen, S., & Nabei, T. (2007). Measuring the effects of oral corrective feedback on L2 knowledge. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 361-377). Oxford: Oxford University Press.

Loewen, S., & Philp, J. (2006). Recasts in the adult English L2 classroom: Characteristics, explicitness, and effectiveness. *Modern Language Journal, 90,* 536-556.

Loewen, S., & Reinders, H. (2011). *Key concepts in second language acquisition*. New York, NY: Palgrave Macmillan.

Long, M. H. (1981). Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences*, *379*(1), 259-278.

Long, M. H. (1983). Linguistic and conversational adjustments to non-native speakers. *Studies in Second Language Acquisition, 5*(2), 177- 193.

Long, M. H. (1985). A role of instruction in second language acquisition: Task based language teaching. In K. Hyltenstam, & M. Pienemann (Eds.), *Modeling and assessing second language development* (pp. 77-99). Clevedon: Multilingual Matters.

Long, M. H. (1990). The least a second language acquisition theory needs to explain. *TESOL Quarterly*, *24*(4), 649-666.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego, CA: Academic Press.

Long, M. H. (2007). *Problems in SLA.* Mahwah, NJ: Lawrence Erlbaum Associates.

Long, M. H. (2015). *Second language acquisition and task-based language teaching.* Oxford: John Wiley & Sons.

Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *Modern Language Journal, 82*(3), 357-371.

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research and practice. In C. Doughty, & J. Williams (Eds.), *Focus on form in second language acquisition* (pp. 15-41). Cambridge: Cambridge University Press.

Lord, G. (2005). (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania, 88*(3), 557-567.

Lyster, R. (1998a). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition*, *20*(1), 51-81.

Lyster, R. (1998b). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, *48*(2), 183-218.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition, 26*(3), 399-432.

Lyster, R., & Izquierdo, J. (2009). Prompts versus recasts in dyadic interaction. *Language Learning*, *59*(2), 453-498.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 20*(1), 37-66.

Lyster, R., & Ranta, L. (2013). Counterpoint piece: The case for variety in corrective feedback research. *Studies in Second Language Acquisition, 35*(1), 167-184.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32*, 265-302.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1-40.

MacDonald, S. (2002). Pronunciation: Views and practices of reluctant teachers. *Prospect, 17*(3), 3-18.

Mackey, A. (2002). Beyond production: Learners' perceptions about interactional processes. *International Journal of Educational Research*, *37,* 379-394.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics, 27*(3), 405-430.

Mackey, A. (2007). Introduction: The role of conversational interaction in second language acquisition. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 1-26). Oxford: Oxford University Press.

Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford: Oxford University Press.

Mackey, A., Abbuhl, R., & Gass, S. M. (2012). Interactionist approach. In S. M. Gass, & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 7-23). New York, NY: Routledge.

Mackey, A., Al-Khalil, M., Atanassova, G., Hama, M., Logan-Terry, A., & Nakatsukasa, K. (2007). Teachers' intentions and learners' perceptions about corrective feedback in the L2 classroom. *International Journal of Innovation in Language Learning and Teaching*, *1*(1), 129-152.

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design.* New York, NY: Routledge.

Mackey, A., Gass, S. M., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition, 22*(4), 471-497.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407-452). Oxford: Oxford University Press.

Mackey, A., Oliver, R., & Leeman, J. (2003). Interactional input and the incorporation of feedback: An exploration of NS–NNS and NNS–NNS adult and child dyads. *Language Learning, 53*(1), 35-66.

Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *The Modern Language Journal, 82,* 338-356.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback, and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp.181-209). Philadelphia, PA: John Benjamins.

Mackey, A., & Polio, C. (2009). Introduction. In A. Mackey, & C. Polio (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 1-10). New York, NY: Routledge.

Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning, 62*(3), 704-740.

Mackey, A., & Silver, R. E. (2005). Interactional tasks and English L2 learning by immigrant children in Singapore. *System, 33*(2), 239-260.

Major, R. C. (1998). Interlanguage phonetics and phonology: An introduction. *Studies in Second Language Acquisition*, *20*(2), 131-137.

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Major, R. C., & Faudree, M. C. (1996). Markedness universals and the acquisition of voicing contrasts by Korean speakers of English. *Studies in Second Language Acquisition*, *18*(1), 69-90.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53-85.

Mattys, S. L. (2000). The perception of primary and secondary stress in English. *Perception & psychophysics*, *62*(2), 253-265.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953-978.

McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics*, *30*(2), 229-258.

McDonough, K. (2007). Interactional feedback and the emergence of simple past activity verbs in L2 English. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 323-338). Oxford: Oxford University Press.

McDonough, K., Crowther, D., Kielstra, P., & Trofimovich, P. (2015). Exploring the potential relationship between eye gaze and English L2 speakers’ responses to recasts. *Second Language Research*, *31*(4), 563-575.

McDonough, K., & Mackey, A. (2006). Responses to recasts: Repetitions, primed production, and linguistic development. *Language Learning, 56*(4), 693-720.

McDonough, K., & Mackey, A. (2008). Syntactic priming and ESL question development. *Studies in Second Language Acquisition, 30*(1), 31-47.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? *A British Society of Audiology Cognition in Hearing Special Interest Group 'White Paper', International Journal of Audiology, 53*(7), 433-445.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 621-638.

Mennen, I. (1998). Can language learners ever acquire the intonation of a second language? *Proceedings of the ESCAL Workshop on Speech Technology in Language Learning*, 17-20.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL quarterly*, *25*(3), 481-520.

Munro, M. J. (1993). Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. *Language and Speech*, *36*(1), 39-66.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285-310.

Munro, M. J., Flege, J. E., & MacKay, I. R. (1996). The effects of age of second language learning on the production of English vowels. *Applied Psycholinguistics*, *17*(3), 313-334.

Munro, M., & Mann, V. (2005). Age of immersion as a predictor of foreign accent. *Applied Psycholinguistics*, *26*(3), 311-341.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, *66*(4), 834-871.

Mustafawi, E. (2018). Arabic phonology. In E. Benmamoun, & R. Bassiouney (Eds.), *The Routledge handbook of Arabic linguistics* (pp. 11-31). New York, NY: Routledge.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

Nassaji, H. (2009). Effects of recasts and elicitations in dyadic interaction and the role of feedback explicitness. *Language Learning*, *59*(2), 411-452.

Nguyễn, T. A. T., & Ingram, C. L. J. (2005). Vietnamese acquisition of English word stress. *TESOL Quarterly*, *39*(2), 309-319.

Nguyễn, T. A. T., Ingram, C. L. J., & Pensalfini, J. R. (2008). Prosodic transfer in Vietnamese acquisition of English contrastive stress patterns. *Journal of Phonetics*, *36*(1), 158-190.

Nicholas, H., Lightbown, P. M., & Spada, N. (2001). Recasts as feedback to language learners. *Language Learning*, *51*(4), 719-758.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417-528.

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, *29*(4), 557-581.

Oliver, R. (1995). Negative feedback in child NS-NNS conversation. *Studies in Second Language Acquisition*, *17*(4), 459-481.

Oliver, R. (2000). Age differences in negotiation and feedback in classroom and pairwork. *Language Learning*, *50*(1), 119-151.

Oliver, R. (2009). How young is too young? Investigating negotiation of meaning and feedback in children aged five to seven years. In A. Mackey, & C. Polio (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 135-156). New York, NY: Routledge.

Oliver, R., & Mackey, A. (2003). Interactional context and feedback in child ESL classrooms. *The Modern Language Journal*, *87*(4), 519-533.

Oller, J. W., & Ziahosseiny, S. M. (1970). The contrastive analysis hypothesis and spelling errors. *Language Learning*, *20*(2), 183-189.

Ortega-Llebaria, M. (2006). Phonetic cues to stress and accent in Spanish. In M. Díaz-Campos (Ed.), *Selected Proceedings of the 2nd conference on Laboratory Approaches to Spanish Phonetics and Phonology* (pp. 104-118), Somerville, MA: Cascadilla Proceedings Project.

Ortega-Llebaria, M., & Prieto, P. (2007). Disentangling stress from accent in Spanish: Production patterns of the stress contrast in deaccented syllables. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 155–176). Amsterdam: John Benjamins.

Ortega-Llebaria, M., & Prieto, P. (2010). Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Language and Speech*, *54*(1), 73-97.

Panova, I., & Lyster, R. (2002). Patterns of corrective feedback and uptake in an adult ESL classroom. *TESOL Quarterly*, *36*(4), 573-595.

Parlak, Ö. (2010). Does pronunciation instruction promote intelligibility and comprehensibility? *SPLIS News*, 7(2), 1-5.

Parlak, Ö, & Ziegler, N. (2017). The impact of recasts on the development of primary stress in a synchronous computer-mediated environment. *Studies in Second Language Acquisition*, *39*(2), 257-285.

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of Speech, Language, and Hearing Research*, *54*(4), 1051-1059.

Peirce, J. W. (2009) Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2,* (10). doi:10.3389/neuro.11.010.2008

Philp, J. (2003). Constraints on "noticing the gap": Nonnative speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition, 25,* 99-126.

Philp, J., & Iwashita, N. (2013). Talking, tuning in and noticing: Exploring the benefits of output in task-based peer interaction. *Language Awareness*, *22*(4), 353-370.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, *44*(3), 493-527.

Pica, T. (1996a). The essential role of negotiation in the communicative classroom. *JALT Journal*, *18*(2), 241-268.

Pica, T. (1996b). Second language learning through interaction: Multiple perspectives. *Working Papers in Educational Linguistics*, *12*(1), 1-22.

Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, *11*(1), 63-90.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation). Massachusetts Institute of Technology, MA.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191-215.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655-687.

Plonsky, L. (2015). Introduction. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159-181. New York, NY: Routledge.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*(2), 325-366.

Plonsky, L., & Oswald, F. L. (2016). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, *39*(3), 579-592.

Qualtrics. (2018). Survey tool [Web-based software]. Accessed through https://lancasteruni.eu.qualtrics.com

Quené, H. (2015). *Package 'hqmisc'*. Retrieved from https://cran.r-project.org/web/packages/hqmisc/hqmisc.pdf

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413-425.

R Core Team. (2017). R: A language and environment for statistical computing

(Version 3.3.3) [Computer program]. Retrieved from https://www.R-project.org/

Rácz, P. (2012). Operationalising salience: Definite article reduction in the North of England. *English Language & Linguistics*, *16*(1), 57-79.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595-626.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*(2), 299-334.

Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*(1), 93-132.

Révész, A., & Han, Z. (2006). Task content familiarity, task type and efficacy of recasts. *Language Awareness*, *15*(3), 160-179.

Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning*, *64*(3), 615-650.

Richards, M. G. (2016). *Not all word stress errors are created equal: Validating an English word stress error gravity hierarchy* (Unpublished doctoral dissertation). Iowa State University, Ames, IA.

Roach, P. (2000). *English phonetics and phonology: A practical course* (3rd ed.). Cambridge: Cambridge University Press.

Roberts, M. (1995). Awareness and the efficacy of error correction. In R. W. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 163-182). Honolulu, HI: University of Hawai'i Press.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language*

Learning, *45*(2), 283-331.

Robinson, P. (1997). Individual differences and the fundamental similarity of implicit and explicit adult second language learning. *Language Learning*, *47*(1), 45-99.

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 631-678). Malden, MA: Blackwell Publishing.

Roettger, T., & Gordon, M. (2017). Methodological issues in the study of word stress correlates. *Linguistics Vanguard, 3*(1).

Rogers, J. (2017). Awareness and learning under incidental learning conditions. *Language Awareness*, *26*(2), 113-133.

RStudio Team. (2016). RStudio: Integrated development for R. (Version 1.1.383) [Computer program]. Retrieved from https://www.rstudio.com

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar. In J. M. Norris, & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133-162). Philadelphia, PA: John Benjamins Publishing Company.

Saito, K. (in press). Corrective feedback and the development of L2 pronunciation. In H. Nassaji, & E. Kartchava (Eds.), *The Cambridge handbook of corrective feedback in language learning and teaching*. Cambridge: Cambridge University Press.

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly*, *46*(4), 842-854.

Saito, K. (2013a). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English /ɹ/ to adult

Japanese learners. *Language Learning*, *63*(3), 499-529.

Saito, K. (2013b). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition*, *35*(1), 1-29.

Saito, K. (2015a). Communicative focus on second language phonetic form: Teaching Japanese learners to perceive and produce English /ɹ/ without explicit instruction. *Applied Psycholinguistics*, *36*(2), 377-409.

Saito, K. (2015b). Variables affecting the effects of recasts on L2 pronunciation development. *Language Teaching Research*, *19*(3), 276-300.

Saito, K. (2017). Effects of sound, vocabulary, and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. *Language Learning*, *67*(3), 665-693.

Saito, K., & Lyster, R. (2012a). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning*, *62*, 595-633.

Saito, K., & Lyster, R. (2012b). Investigating the pedagogical potential of recasts for L2 vowel acquisition. *TESOL Quarterly*, *46*(2), 387-398.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217-240.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439-462.

Saito, K., & Wu, X. (2014). Communicative focus on form and second language suprasegmental learning: Teaching Cantonese learners to perceive Mandarin tones. *Studies in Second Language Acquisition*, *36*(4), 647-680.

Salient (n.d.). In *Merriam-Webster Dictionary online*. Retrieved from https://www.merriam-webster.com/dictionary/salient

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning.* New York, NY: Palgrave MacMillan.

Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning & Technology*, *13*(1), 96-120.

Saxton, M. (1997). The contrast theory of negative input. *Journal of Child Language, 24*(1), 139-161.

Saxton, M. (2005). 'Recast' in a new light: Insights for practice from typical language studies. *Child Language Teaching and Therapy*, *21*(1), 23-38.

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly, 40*(4), 715-738.

Schachter, J. (1974). An error in error analysis. *Language Learning*, *24*(2), 205-214.

Schachter, J. (1984). A universal input condition. In W. E. Rutherford (Ed.), *Universals and second language acquisition* (pp. 167-183). Philadelphia, PA: John Benjamins.

Schachter, J. (1986). Three approaches to the study of input. *Language Learning, 36*(2) 211-225.

Schachter, J. (1991). Corrective feedback in historical perspective. *Second Language Research*, *7*(2), 89-102.

Schielzeth, H. & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology, 20*, 416-420.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129-158.

Schmidt, R. W. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*(4), 357-385.

Schmidt, R. W. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206-226.

Schmidt, R. W. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *Consciousness in Second Language Learning*, *11*, 237-326.

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 1-32). Cambridge, UK: Cambridge University Press.

Schmidt, R. W., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237-326). Rowley, MA: Newbury House.

Scovel, T. (2000). A critical review of critical period research. *Annual Review of Applied Linguistics, 20*, 213–223.

Seedhouse, P. (2005). "Task" as research construct. *Language Learning*, *55*(3), 533-570.

Seidlhofer, B. (2004). Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics*, *24*, 209-239.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(3), 209-232.

Setter, J. & Sebina, B. (2018). English lexical stress, prominence and rhythm. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 137-153). London: Routledge.

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research, 8*(3), 263-300.

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research, 10*(4), 361-392.

Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 301-322). Oxford: Oxford University Press.

Sheen, Y. (2008). Recasts, language anxiety, modified output, and L2 learning. *Language Learning, 58*(4), 835-874.

Sheen, Y., & Ellis, R. (2011). Corrective feedback in language teaching. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*, (pp. 593-610). New York, NY: Routledge.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass, 3*(2), 621-640.

Skinner, B. (1957). *Verbal behavior*. New York, NY: Appleton Century Crofts.

Sluijter, A. M., & Van Heuven, V. J. (1996a). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America*, *100*(4), 2471-2485.

Sluijter, A. M., & van Heuven, V. J. (1996b). Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proceedings of the Fourth International Conference on Spoken Language Processing, 2,* 630-633.

Sluijter, A. M., Van Heuven, V. J., & Pacilly, J. J. (1997). Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America*, *101*(1), 503-513.

Smith, M. S. (1991). Speaking to many minds. *Second Language Research, 7*, 118-132.

Smith, M. S. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, *15*(2), 165-179.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language, 45*(3), 412-432.

Spada, N., & Lightbown, P. M. (2009). Interaction research in second/foreign language classrooms. In A. Mackey, & C. Polio (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 157-175). New York, NY: Routledge.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1-19.

Swain, M. (1985) Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass, & C. Madden (Eds.), *Input and second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook, & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honor of H. G. Widdowson* (pp. 125-144). Oxford: Oxford University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471-483). Mahwah, NJ: Lawrence Erlbaum Associates.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371-391.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, *82*(3), 320-337.

Swain, M., & Lapkin, S. (2002). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research*, *37*, 285-304.

Taylor, S., & Wales, R. (1987). Primitive mechanisms of accent perception. *Journal of Phonetics, 15*(3), 235-246.

Taft, M. (2002). Orthographic processing of polysyllabic words by native and nonnative English speakers. *Brain and Language, 81*, 532-544.

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal, 28*(3), 744-765.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*(3), 326-344.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, *16*(2), 183-203.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569-613.

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of* *empirical studies* (pp. 171-195). Oxford: Oxford University Press.

Trofimovich, P. & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28,*1-30.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905-916.

Trofimovich, P., McDonough, K., & Foote, J. A. (2014). Interactive alignment of multisyllabic stress patterns in a second language classroom. *TESOL Quarterly*, *48*(4), 815-832.

Turk, A. E., & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America*, *99*(6), 3782-3790.

Ullakonoja, R. (2007). Comparison of pitch range in Finnish (L1) and Russian (L2). In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, 1701-1704.

Van den Branden, K. (2006). Introduction: Task-based language teaching in a nutshell. In K. Van den Branden (Ed.), *Task-based language education: From theory to practice* (pp. 1-16). Cambridge: Cambridge University Press.

van der Hulst, H. (2003). Stress and accent. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (pp. 246-254). London: Nature Publishing Group.

van der Hulst, H. (2012). Deconstructing stress. *Lingua*, *122,* 1494–1521.

van der Hulst, H. (2014). The study of word accent and stress: Past, present, and future. In H. van der Hulst (Ed.), *Word stress: Theoretical and typological issues* (pp. 3-55). Cambridge: Cambridge University Press.

van Donselaar, W., Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in recognition. *Quarterly Journal of Experimental Psychology, 58A*, 251-273.

VanPatten, B., & Benati, A. (2015). *Key terms in second language acquisition*. London: Bloomsbury Academic.

VanPatten, B., & Williams, J. (2015). Theories in second language acquisition: An introduction. *Introduction: Nature of theories* (pp. 1-16). New York, NY: Routledge.

Varonis, E. M., & Gass, S. (1985). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics*, *6*(1), 71-90.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Acoustical Society of America*, *113*(2), 1033-1043.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*(4), 539-552.

Wardhaugh, R. (1970). The contrastive analysis hypothesis. *TESOL Quarterly*, 123-130.

Watson, J. C. (2007). *The phonology and morphology of Arabic.* New York, NY: Oxford University Press.

West, B. T., Welch, K. B., & Galecki, A. T. (2015). *Linear mixed models: A practical guide using statistical software* (2nd ed.). Boca Raton, FL; CRC Press.

White, J. (1998). Getting the learners' attention: A typographical input enhancement study. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 85-113). Cambridge: Cambridge University Press.

White, L. (1991). Adverb placement in second language acquisition: Some effects of positive and negative evidence in the classroom. *Interlanguage Studies Bulletin*, *7*(2), 133-161.

White, L. (1989). *Universal grammar and second language acquisition*. Amsterdam: John Benjamins.

Whitman, R. L., & Jackson, K. L. (1972). The unpredictability of contrastive analysis. *Language Learning*, *22*(1), 29-41.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis.* Switzerland: Springer.

Williams, J. (1999). Learner-generated attention to form. *Language Learning*, *49*(4), 583-625.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, *47*(4), 762-789.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling, *Journal of Language Evolution*, *1*(1), 7-18. doi.org/10.1093/jole/lzv003

Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, *55*(3), 486-507.

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, *38*(3), 329-336.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics*, *34*(3), 344-368.

Yilmaz, Y. (2016). The linguistic environment, interaction and negative feedback. *Brill Research Perspectives on Multilingualism and Second Language Acquisition, 1*, 45-86.

Young-Scholten, M. (1998). Second language syllable simplification: Deviant development or deviant input? In J. Allan, & J. Leather (Eds), *New Sounds 97* (pp. 351-360). Amsterdam: University of Amsterdam.

Young-Scholten, M., & Langer, M. (2015). The role of orthographic input in second language German: Evidence from naturalistic adult learners' production. *Applied Psycholinguistics, 36*, 93-114.

Yoshida, R. (2008). Teachers' choice and learners' preference of corrective feedback types. *Language Awareness*, *17*(1), 78-93.

Zampini, M. L. (1998). The relationship between the production and perception of L2 Spanish stops. *Texas Papers in Foreign Language Education, 3*(3), 85-100.

Zellers, M. (2017). Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish. *Language and Speech*, *60*(3), 454-478.

Zhang, Y., & Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics, 38*(2), 260-271.

Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *The Journal of the Acoustical Society of America*, *123*(6), 4498-4513.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 69-84.

Zimmerer, F., Jügler, J., Andreeva, B., Möbius, B., & Trouvain, J. (2014). Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers. *Proceedings of the 7th International Conference of Speech Prosody,* 1037-1041.

Zsiga, E. C. (2013). *The sounds of language: An introduction to phonetics and phonology*. Oxford: John Wiley & Sons.

Zuraiq, W., & Sereno, J. A. (2007). English lexical stress cues in native English and non-native Arabic speakers. *Proceedings of the 16th International Congress of Phonetic Science,* 829-832.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, *1*(1), 3-14.

APPENDICES

Appendix A: Tasks

Appendix B: Background survey for learners

Appendix C: Background survey for listeners

Appendix D: Exit survey for learners

Appendix E: Distribution of data and model residuals

Appendix F: Linear mixed-effects analysis results

Appendix G: Residual plots

Appendix A

Tasks

Task 1a: Below there are 10 incomplete sentences about language teachers and language classrooms. Complete the sentences based on your opinion. There are no right or wrong answers.

Student A
1. Language teachers need to consider their students' interests when creating lessons because ______________________________
2. To encourage students who have a negative perception of language learning, teachers can ______________________________
3. Teachers should be consistent when grading exams because ______________________________
4. An important part of writing assignments is revision because ______________________________
5. When teachers are responsive to student needs, students will ______________________________
6. If a teacher has a suspicion that a student is cheating, the teacher needs to ______________________________
7. A dynamic classroom helps students to ______________________________
8. When there is confusion about the meaning of a word, the teacher ______________________________
9. When teachers develop a good relationship with their students, ______________________________
10. When there is a contention between two students, the teacher should ______________________________

Appendix A (continued)

Task 1b: Below there are 10 incomplete sentences about language teachers and language classrooms. Complete the sentences based on your opinion. There are no right or wrong answers.

Student B
1. If students struggle with sentence formation and grammar, ______________________________ ______________________________
2. In order to diminish issues related to language anxiety, a teacher can ______________________ ______________________________
3. When writing papers, students need to follow the convention of using formal language because______________________________
4. Students need to be persistent when they face difficulties because _________________________ ______________________________
5. Having compassion makes a teacher ______________________________ ______________________________
6. The position of student desks is important because______________________________ ______________________________
7. The best companion of a language learner is ______________________________ ______________________________
8. The ideal duration for a language class should be around ______________________________ ______________________________
9. An important condition that helps language learning is ______________________________ ______________________________
10. If there is a consensus among students that the teacher is not a fair grader, __________________ ______________________________

Appendix A (continued)

Task 3a: With your partner, you will interview a candidate who has applied for a teaching job to teach at the Academic Bridge Program at AUS. Take turns with your partner and interview the candidate by asking the questions below. Take brief notes so that you can remember his answers after the interview.

Student A
1) Do you consider students' interests when creating lessons?
2) How do you encourage students who have a negative perception of language learning?
3) How do you make sure you are consistent when grading exams?
4) How do you teach revision when working on writing assignments?
5) Are you responsive to student needs?
6) When you have a suspicion that a student is cheating, what do you do?
7) How do you create a dynamic classroom?
8) When there is confusion about the meaning of a word, how do you explain it?
9) How do you develop good relationship with your students?
10) When there is a contention between two students, what do you do?

Appendix A (continued)

Task 3b: With your partner, you will interview a candidate who has applied for a teaching job to teach at the Academic Bridge Program at AUS. Take turns with your partner and interview the candidate by asking the questions below. Take brief notes so that you can remember his answers after the interview.

Student B
1) What do you do when your students struggle with sentence formation and grammar?
2) What can be done to diminish issues related to language anxiety?
3) How do you teach your students to use the convention of using formal language for their writing assignments?
4) How do you motivate students to be persistent when they face difficulties?
5) Does having compassion make a teacher better at their job?
6) Is the position of student desks important?
7) What is the best companion of a language learner?
8) What is the ideal duration for a language class?
9) What is an important condition that helps language learning?
10) What should a teacher do if there is a consensus among students that he/she is not a fair grader?

Appendix A (continued)

Task 4a: Below are 10 incomplete sentences about the candidate's responses to the interview questions. Complete the sentences based on the answers provided by the candidate during the interview. (Time: 10 minutes)

Student A
1. He says it is important to consider students' interests because ____________________________ ____
2. When he sees a student with a negative perception of language learning, ___________________ ____
3. To be consistent when grading exams, he ____ ____
4. He teaches revision by ____ ____
5. He says he is responsive to student needs because _____________________________________ ____
6. When he has a suspicion that a student is cheating, _____________________________________ ____
7. To create a dynamic classroom, he ____ ____
8. When there is a confusion about the meaning of a word, he ______________________________ ____
9. In order to develop a good relationship with his students, he ____________________________ ____
10. When there is a contention between two students, he __________________________________ ____

Appendix A (continued)

Task 4b: Below are 10 incomplete sentences about the candidate's responses to the interview questions. Complete the sentences based on the answers provided by the candidate during the interview. (Time: 10 minutes)

Student B
1. When students struggle with sentence formation and grammar, he ________________________ ____
2. In order to diminish issues related to language anxiety, he ________________________________ ____
3. To teach the convention of using formal language, he ___________________________________ ____
4. He motivates students to be persistent by ____ ____
5. He thinks that having compassion makes a teacher ______________________________________ ____
6. He says the position of student desks ____ ____
7. He believes that the best companion of a language learner is _____________________________ ____
8. He thinks the ideal duration for a language class should be _____________________________ ____
9. He says an important condition that helps language learning is ___________________________ ____
10. If there is a consensus among students that the teacher is not a fair grader, he thinks ________ ____

Appendix B

Background Survey for Learners

qualtrics

INSTRUCTIONS

The information from this survey will not be used or distributed outside the current research study. To protect your privacy, a code number will be used instead of your name. Your responses will not be shared with any third party. Remember that your responses are voluntary. Please answer each question to the best of your ability.

Before you continue, please enter the participant code given to you by the researcher.

What is your gender?

- Male
- Female

Are you left-handed or right-handed?

- Left-handed
- Right-handed

How old are you?

Do you have any speech or hearing related problems that you are aware of?

- No
- Yes (Please explain briefly)

Appendix B (continued)

Which Arabic dialect do you speak?

- ◯ Egyptian
- ◯ Gulf
- ◯ Hijazi
- ◯ Iraqi
- ◯ Levantine
- ◯ Najdi
- ◯ North African
- ◯ Sudanese
- ◯ Yemeni

Other than Arabic and English, what languages do you speak fluently?

- ◯ I do not speak any other language fluently.
- ◯ I speak the language(s) below fluently:

Have you lived in an English-speaking country (e.g., the United States, the United Kingdom, Canada, Australia)?

- ◯ No
- ◯ Yes (Please mention where and for how long)

How long have you lived in the United Arab Emirates?

How old were you when you started learning English?

How many years of English instruction have you received?

Appendix B (continued)

Have you taken pronunciation classes before?

- No
- Yes (Please mention where and for how long)

How would you divide the percentage of your use of English among the three situations listed below?

	0	10	20	30	40	50	60	70	80	90	100	
Talking to non-native speakers												0
Talking to native speakers												0
Talking to family when at home												0
Total:												0

Are you a Bridge Program student or an undergraduate student?

- Bridge Program ELP 100
- Bridge Program ELP 200
- Bridge Program ELP 250
- WRI 001
- WRI 101

What is your latest IELTS or TOEFL score?

- IELTS
- TOEFL

When did you take the IELTS or TOEFL?

Appendix C

Background Survey for Listeners

qualtrics

INSTRUCTIONS

The information from this survey will not be used or distributed outside the current research study. To protect your privacy, a code number will be used instead of your name. Your responses will not be shared with any third party. Remember that your responses are voluntary. Please answer each question to the best of your ability.

Before you continue, please enter the participant code given to you by the researcher.

What is your gender?

- Male
- Female

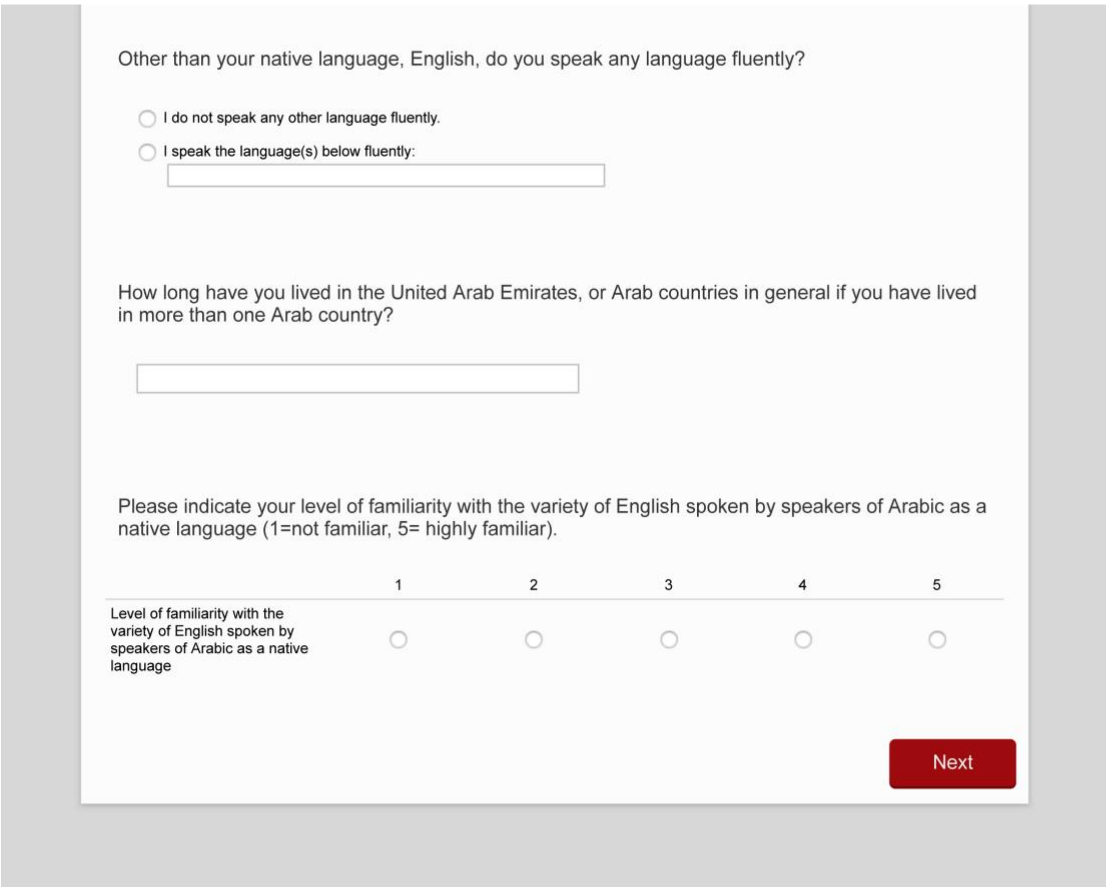
How old are you? (Please enter a number)

Do you have any speech or hearing related problems that you are aware of?

- No
- Yes (Please explain briefly)

Which dialect of English do you speak?

Appendix C (continued)



Other than your native language, English, do you speak any language fluently?

- I do not speak any other language fluently.
- I speak the language(s) below fluently:

How long have you lived in the United Arab Emirates, or Arab countries in general if you have lived in more than one Arab country?

Please indicate your level of familiarity with the variety of English spoken by speakers of Arabic as a native language (1=not familiar, 5= highly familiar).

	1	2	3	4	5
Level of familiarity with the variety of English spoken by speakers of Arabic as a native language	○	○	○	○	○

Next

Appendix D

Exit Survey for Learners

qualtrics

INSTRUCTIONS

The information from this survey will not be used or distributed outside the current research study. To protect your privacy, a code number will be used instead of your name. Your responses are voluntary and will not be shared with any third party. Please answer each question to the best of your ability.

Before you continue, please enter the participant code given to you by the researcher.

Please indicate how much you agree or disagree with the following statements.

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
During the interview task, I focused on word meaning.	○	○	○	○	○
During the interview task, I focused on grammar.	○	○	○	○	○
During the interview task, I focused on pronunciation.	○	○	○	○	○
The interviewee gave me feedback when I used the wrong word.	○	○	○	○	○
The interviewee gave me feedback when I made grammar mistakes.	○	○	○	○	○
The interviewee gave me feedback when I made pronunciation mistakes.	○	○	○	○	○

Please rank the following language features <u>based on the attention you gave</u> during the interview task (1= highest, 3=lowest).

	1	2	3
Grammar	○	○	○
Pronunciation	○	○	○
Word meaning	○	○	○

Appendix D (continued)

Did the interviewee correct your English during the interview task? If yes, what did he correct and how did you react? Please explain.

How much attention (if any) do you feel you paid to pronunciation during the interview task? Can you give any examples?

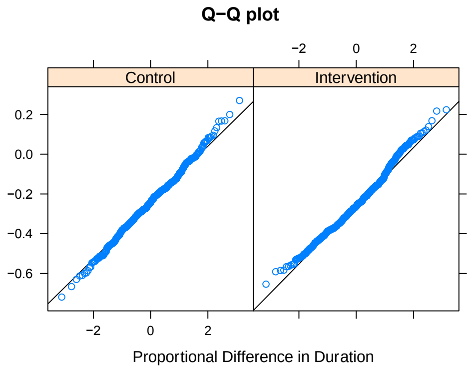
How much attention (if any) do you feel you paid to word meaning during the interview task? Can you give any examples?

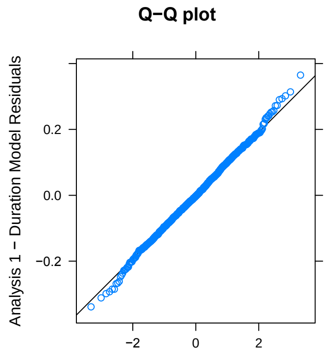
How much attention (if any) do you feel you paid to grammar during the interview task? Can you give any examples?

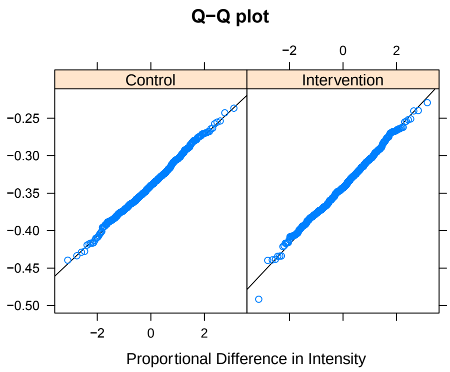
Next

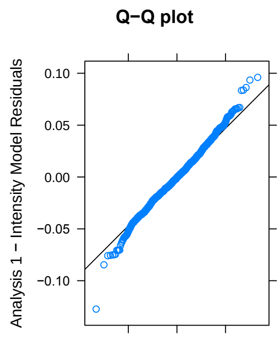
Appendix E

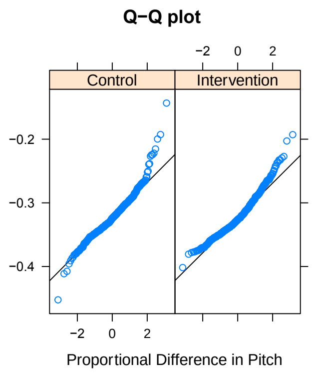
Distribution of Data and Model Residuals: Analysis 1

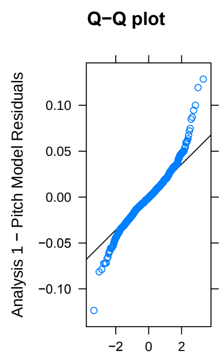






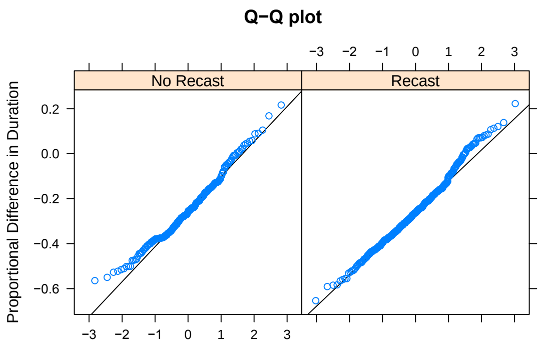


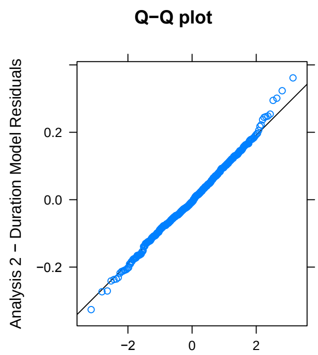


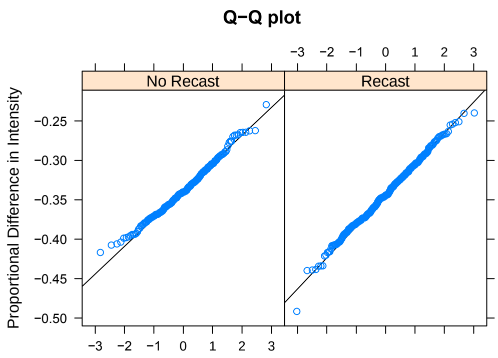


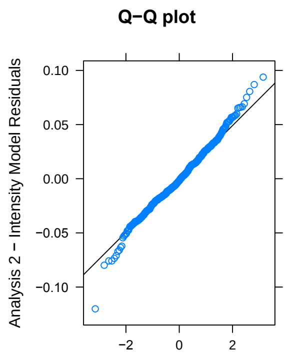
Appendix E (continued)

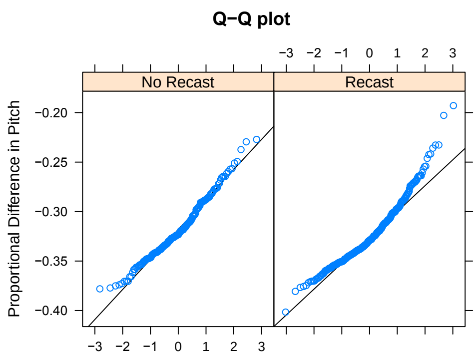
Distribution of Data and Model Residuals: Analysis 2

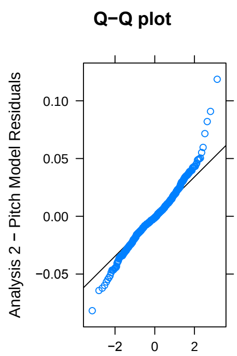












Appendix F

Linear Mixed-Effects Analysis Results: Duration Measures, Analysis 1

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## dur_rd ~ condition * time + (1 | learner) + (0 + time | learner) +
##     (1 | word) + (0 + time | word)
##    Data: data
##
## REML criterion at convergence: -1777.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3227 -0.6357 -0.0318  0.6330  3.5809
##
## Random effects:
##  Groups    Name         Variance  Std.Dev.  Corr
##  learner   (Intercept)  4.724e-15 6.873e-08
##  learner.1 timePretest  1.360e-03 3.688e-02
##            timePosttest 1.076e-03 3.280e-02 1.00
##  word      (Intercept)  5.030e-11 7.092e-06
##  word.1    timePretest  1.060e-02 1.029e-01
##            timePosttest 1.169e-02 1.081e-01 1.00
##  Residual               1.039e-02 1.019e-01
## Number of obs: 1130, groups:  learner, 68; word, 20
##
## Fixed effects:
##                                Estimate Std. Error         df t value
## (Intercept)                   -0.276083   0.024554  22.800000 -11.244
## conditionControl               0.025530   0.012480  67.600000   2.046
## timePosttest                   0.036315   0.008376 443.600000   4.336
## conditionControl:timePosttest  -0.036880   0.012204 804.400000  -3.022
##                               Pr(>|t|)
## (Intercept)                   8.7e-11 ***
## conditionControl              0.04468 *
## timePosttest                  1.8e-05 ***
## conditionControl:timePosttest  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix F (continued)

Linear Mixed-Effects Analysis Results: Intensity Measures, Analysis 1

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## int_rd ~ condition * time + (1 | learner) + (0 + time | learner) +
##     (1 | word) + (0 + time | word)
##    Data: data
##
## REML criterion at convergence: -4832.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.8032 -0.6021 -0.0183  0.5926  3.6223
##
## Random effects:
##  Groups    Name         Variance  Std.Dev. Corr
##  learner   (Intercept)  0.000e+00 0.000000
##  learner.1 timePretest  4.851e-05 0.006965
##            timePosttest 6.073e-05 0.007793 1.00
##  word      (Intercept)  0.000e+00 0.000000
##  word.1    timePretest  5.543e-04 0.023543
##            timePosttest 4.813e-04 0.021939 1.00
##  Residual               7.032e-04 0.026518
## Number of obs: 1130, groups:  learner, 68; word, 20
##
## Fixed effects:
##                                Estimate Std. Error         df t value
## (Intercept)                   -0.342823   0.005607  22.300000 -61.137
## conditionControl               0.004229   0.002819  79.600000   1.500
## timePosttest                   0.002298   0.002185 384.100000   1.052
## conditionControl:timePosttest  -0.002329   0.003171 878.500000  -0.735
##                               Pr(>|t|)
## (Intercept)                    <2e-16 ***
## conditionControl                0.137
## timePosttest                    0.294
## conditionControl:timePosttest    0.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix F (continued)

Linear Mixed-Effects Analysis Results: Pitch Measures, Analysis 1

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula: pitst_rd ~ condition * time + (1 + time | learner) + (1 + time |
##     word)
##    Data: data
##
## REML criterion at convergence: -5061.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.3020 -0.5214 -0.0262  0.5131  5.5123
##
## Random effects:
##  Groups   Name         Variance  Std.Dev.  Corr
##  learner  (Intercept)  1.680e-04 0.0129608
##           timePosttest 2.479e-05 0.0049794 1.00
##  word     (Intercept)  1.270e-04 0.0112700
##           timePosttest 9.489e-07 0.0009741 -1.00
##  Residual              5.428e-04 0.0232977
## Number of obs: 1130, groups:  learner, 68; word, 20
##
## Fixed effects:
##                                Estimate Std. Error         df t value
## (Intercept)                   -0.325501   0.003586  45.220000 -90.774
## conditionControl               0.001402   0.003723  68.380000   0.377
## timePosttest                   0.006803   0.002076 164.140000   3.277
## conditionControl:timePosttest  -0.004723   0.003034 183.840000  -1.557
##                               Pr(>|t|)
## (Intercept)                   < 2e-16 ***
## conditionControl              0.70766
## timePosttest                  0.00128 **
## conditionControl:timePosttest  0.12117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix F (continued)

Linear Mixed-Effects Analysis Results: Duration Measures, Analysis 2

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## dur_rd ~ feedback * time + (1 | learner) + (0 + time | learner) +
##     (1 | word) + (0 + time | word)
##    Data: dataintv
##
## REML criterion at convergence: -937.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2254 -0.6263 -0.0687  0.6438  3.5750
##
## Random effects:
##  Groups    Name         Variance  Std.Dev.  Corr
##  learner   (Intercept)  0.000e+00 0.000e+00
##  learner.1 timePretest  1.008e-03 3.176e-02
##            timePosttest 8.513e-04 2.918e-02 1.00
##  word      (Intercept)  7.741e-12 2.782e-06
##  word.1    timePretest  9.183e-03 9.583e-02
##            timePosttest 1.210e-02 1.100e-01 1.00
##  Residual               1.023e-02 1.011e-01
## Number of obs: 608, groups:  learner, 36; word, 20
##
## Fixed effects:
##                                Estimate Std. Error        df t value
## (Intercept)                    -0.28856    0.02327  22.90000 -12.399
## feedbackNo Recast               0.03634    0.01285 549.50000   2.827
## timePosttest                    0.05539    0.01063 133.70000   5.209
## feedbackNo Recast:timePosttest  -0.05703    0.01726 551.20000  -3.305
##                               Pr(>|t|)
## (Intercept)                   1.25e-11 ***
## feedbackNo Recast              0.00486 **
## timePosttest                  7.00e-07 ***
## feedbackNo Recast:timePosttest  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix F (continued)

Linear Mixed-Effects Analysis Results: Intensity Measures, Analysis 2

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## int_rd ~ feedback * time + (1 | learner) + (0 + time | learner) +
##     (1 | word) + (0 + time | word)
##    Data: dataintv
##
## REML criterion at convergence: -2525.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3679 -0.6115 -0.0633  0.5972  3.4120
##
## Random effects:
##  Groups    Name         Variance  Std.Dev.  Corr
##  learner   (Intercept)  4.368e-17 6.609e-09
##  learner.1 timePretest  2.128e-05 4.614e-03
##            timePosttest 6.391e-05 7.994e-03 1.00
##  word      (Intercept)  4.598e-17 6.781e-09
##  word.1    timePretest  6.632e-04 2.575e-02
##            timePosttest 4.446e-04 2.108e-02 1.00
##  Residual               7.564e-04 2.750e-02
## Number of obs: 608, groups:  learner, 36; word, 20
##
## Fixed effects:
##                                 Estimate Std. Error         df t value
## (Intercept)                    -0.344321   0.006142  21.300000 -56.057
## feedbackNo Recast               0.004299   0.003472 539.800000   1.238
## timePosttest                    0.001777   0.003006  81.000000   0.591
## feedbackNo Recast:timePosttest  0.001427   0.004705 549.100000   0.303
##                               Pr(>|t|)
## (Intercept)                     <2e-16 ***
## feedbackNo Recast                0.216
## timePosttest                     0.556
## feedbackNo Recast:timePosttest   0.762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

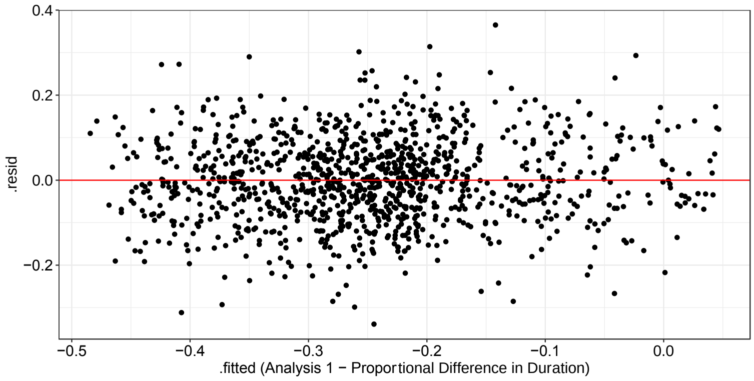
Appendix F (continued)

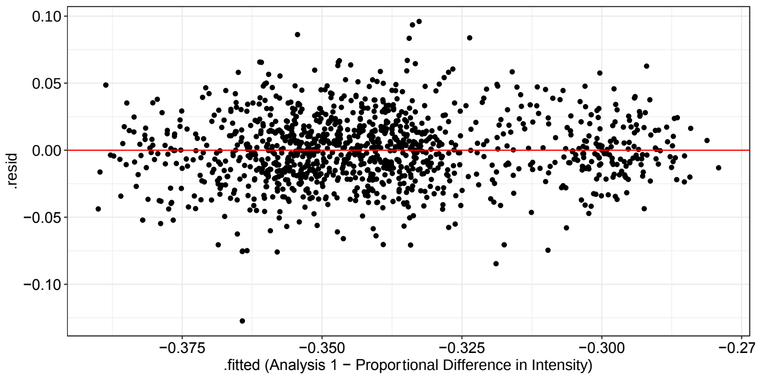
Linear Mixed-Effects Analysis Results: Pitch Measures, Analysis 2

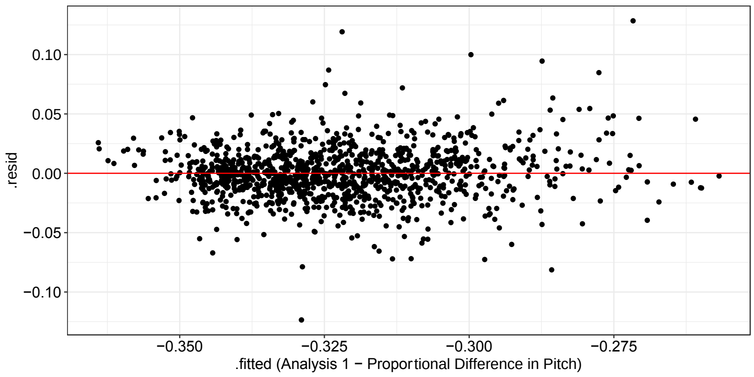
```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula: pitst_rd ~ feedback * time + (1 + time | learner) + (1 + time |
##     word)
##    Data: dataintv
##
## REML criterion at convergence: -2740.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6468 -0.5286 -0.0534  0.5054  5.2934
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.  Corr
##  learner  (Intercept)  1.259e-04 0.0112214
##           timePosttest 3.863e-05 0.0062157 1.00
##  word     (Intercept)  1.276e-04 0.0112981
##           timePosttest 5.149e-07 0.0007176 1.00
##  Residual              5.024e-04 0.0224150
## Number of obs: 608, groups:  learner, 36; word, 20
##
## Fixed effects:
##                                 Estimate Std. Error         df t value
## (Intercept)                    -0.327816   0.003542  34.600000 -92.555
## feedbackNo Recast               0.006835   0.002839 542.200000   2.407
## timePosttest                    0.009052   0.002484 116.600000   3.643
## feedbackNo Recast:timePosttest  -0.006531   0.003837 549.700000  -1.702
##                                Pr(>|t|)
## (Intercept)                     < 2e-16 ***
## feedbackNo Recast              0.016396 *
## timePosttest                   0.000403 ***
## feedbackNo Recast:timePosttest 0.089323 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix G

Residual Plots: Analysis 1







Appendix G (continued)

Residual Plots: Analysis 2