NORTHWESTERN UNIVERSITY

The Role of Linguistic Experience in the Production and Perception of Probabilistic Reduction

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

by

Erin Gustafson

EVANSTON, ILLINOIS

December 2016

ABSTRACT

In this dissertation, we present three empirical studies investigating the role of linguistic experience in the processing of probabilistic information during speech production, speech perception, and across modalities. In all studies, we focus on a particular type of probabilistic information related to the probability of a word in a discourse (i.e., whether a word is discourse-given with high probability or discourse-new with low-probability). Study 1 examines how variation in discourse-dependent probability shapes the phonetic properties of content and function words during production. We test both first language (L1) and second language (L2) speakers in order to better understand how linguistic experience impacts the processing of probabilistic information. Differences between these groups in the production of content vs. function words provide insights into the mechanism underlying the influence of probabilistic information on production processing. In Study 2, we ask whether linguistic experience impacts listeners' ability to use probabilistic information (i.e., the reduction associated with high vs. low discourse-dependent probability) as a predictive cue during speech perception. Prediction can pose a challenge for L2 listeners, who may lack sufficient experience with the structures necessary to engage predictive processing. Differences between groups raise questions about the mechanisms underlying prediction during speech perception. Finally, Study 3 investigates the coupling of production and perception in terms of how probabilistic information influences processing. Relations among individual differences between L2 participants in the first two studies shed light on similarities across modalities in how probabilistic information influences production and perception processing. Together, the results of these three studies provide a sketch of how probabilistic information influences speech behavior across individuals with varying levels of linguistic experience.

ACKNOWLEDGEMENTS

I cannot adequately express my gratitude to all of the people in my life who have supported me along this journey to get my PhD.

To my advisor and mentor Matt Goldrick, thank you for inspiring my love of psycholinguistics and data analysis. Your faith in my abilities and willingness to push me out of my comfort zone have allowed me to mature as a linguist, scientist, and thinker. Our work together over the years has taught me to think and write with precision, to build strong and thorough arguments, and to approach data analysis in innovative ways. I will be forever grateful to you for the time you devoted to my mentorship.

My academic and intellectual development over these past five years has also been heavily shaped by a member of my dissertation committee and QP2 advisor, Ann Bradlow. As a member of your lab, I was challenged to step back and consider my research from a different point of view. Your perspectives have demonstrably increased the quality of my work, and I am so thankful that I had the opportunity to work closely with you on my QP2. To my other committee member, Klinton Bicknell, thank you for introducing me to computational approaches to linguistics research. Your perspectives complement Matt and Ann's so well, and my dissertation committee is stronger because you are a part of it.

I never would have survived grad school if not for my wonderful, supportive, and brilliant peers. Thanks to members of SoundLab and the Speech Communication Research Group for your honest and constructive feedback on my work. I especially thank Tommy Denby and Emily Cibelli, who took the time to read and comment on Chapters 2-4 of this dissertation. An enormous thank you goes to my cohort, past and present: Alex, Angela, Chelsea, Jeremy, Kristin, Peter, and Sveltin. Our first year – and every year, really – would have been impossible

without you guys here to bounce ideas around with and goof around with. Thanks to the Angelas (Cooper and Fink) for our many girls' nights to watch movies, enjoy bubbles, and gab incessantly. I value your friendship more than you know. Finally, a special thank you to Julie Matsubara for helping me get through this last push to finish writing my dissertation. You kept me accountable and kept me motivated; I couldn't have done it without you!

I thank my family for their support in all things, but especially their unfailing support during this crazy journey. To my mom, Connie Laufersky, our weekly phone calls keep me grounded. Your intellectual curiosity and intelligence are a huge inspiration to me. You are exceptional person and the best mother. I am so lucky. Thank you for everything. To my dad, John Gustafson, thank you for all of the advice and perspective that you have offered over these last five years. I can honestly say that I could not have finished this impossible task without you right down the street to give me some much needed distractions and feed me dinner. I feel like this experience has brought us closer, and I am so lucky to have you as my dad. Finally, to my best friend and little seester, Megan, thank you for listening to me complain on a daily basis. Our constant contact keeps me sane. Thanks for always being there for me.

And last but certainly not least… thank you to my boyfriend, Patrick Koffler. Thank you for being by my side for the last three years, constantly reminding me that I can do this. Thank you for helping me balance work and life, for ordering Thai food, for weekend mornings at the movie theater, for soccer Saturdays and football Sundays, for reminding me that I should probably be writing, for all of the laughs, for your honesty, and for loving me no matter what. You help me get through each and every day. Thank you from the bottom of my heart. Love you always.

TABLE OF CONTENTS

ABSTRACT

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

1.1 Introduction

1.2 Study 1: Linguistic Experience Influences the Processing of Probabilistic Information during Function but not Content Word Production

1.3 Study 2: Linguistic Experience Impacts the Time-Course of Prediction during Speech Perception

1.4 Study 3: Similarities in How Probabilistic Processing Influences Production and Perception: Cross-Task Transfer

1.5 Conclusions

CHAPTER 2

2.1 Introduction to Study 1

2.1.1 The Processing of Probabilistic Information in Speech Production

2.1.2 Processing of Content vs. Function Words

2.1.3 Effects of Variation in Content Word Probabilities

2.1.3.1 L1 speech

2.1.3.2 L2 speech

2.1.3.3 Study goal 1: L1 vs. L2 differences in the effects of content word probabilities in a more demanding task

2.1.4 Effects of Variation in Function Word Probabilities

2.1.4.1 L1 speech

2.1.4.2 L2 speech

2.1.4.3 Study goal 2: L1 vs. L2 differences in the effects of function word probabilities in a more demanding task

2.1.5 The Current Study: Summary

2.2 Method

2.2.1 Participants

2.2.2 Materials and Design

2.2.3 Procedure

2.2.4 Measurement

2.2.5 Analysis

2.2.5.1 Duration measures

2.2.5.2 Disfluencies

2.3 Results

2.3.1 Response Times

2.3.2 Noun Durations

2.3.3 Determiner Vowel Durations

2.3.4 Disfluencies

2.3.5 Post-hoc Analysis: Lexical Frequency

2.4 General Discussion

2.4.1 L2 Processing of Probabilistic Information during Noun Production

2.4.2 L2 Processing of Probabilistic Information during Determiner Production

2.4.3 Probabilistic Information for the Speaker vs. for the Listener

2.5 Conclusions

CHAPTER 3

3.1 Introduction to Study 2

3.1.1 Processing Probabilistic Information in Production

3.1.2 Prediction in L1 Perception

3.1.3 Prediction in L2 Perception

3.1.4 The Current Study

3.2 Experiment 1

3.2.1 Method

3.2.1.1 Participants

3.2.1.2 Materials and design

3.2.1.3 Recordings

3.2.1.4 Procedure

3.2.1.5 Additional tasks

3.2.1.6 Data pre-processing

3.2.1.7 Accuracy analysis

3.2.1.8 Eye movement analysis

3.2.1.9 Eye movement hypotheses and predictions

3.2.2 Results

3.2.2.1 Accuracy

3.2.2.2 Eye movements in the early window

3.2.2.3 Eye movements to the target in the late window

3.2.2.3.1 Fixations overall

3.2.2.3.2 Fixations over time: Rate of looking towards and away from target

3.2.2.3.3 Fixations over time: Maintaining target fixations

3.2.2.4 Eye movements to the competitor in the late window

3.2.2.4.1 Fixations overall

3.2.2.4.2 Fixations over time: Drop-off in looks to competitor

3.2.2.4.3 Fixations over time: Slope and peakedness of competitor fixations

3.2.2.5 Interim discussion

3.3 Experiment 2

3.3.1 Method

3.3.1.1 Participants

3.3.1.2 Data pre-processing

3.3.1.3 Eye movement analysis

3.3.2 Results

3.3.2.1 Eye movements in the early window

3.3.2.2 Eye movements to the target in the late window

3.3.2.3 Eye movements in the competitor in the late window

3.3.2.3.1 Fixations overall

3.3.2.3.2 Fixations over time

3.3.2.4 Interim discussion

3.4 General Discussion

3.4.1 Prediction by L1 and L2 Listeners from Target-Concurrent Information

3.4.2 Lack of Prediction from Contextual Information

3.4.3 Mechanisms Underlying Prediction

3.5 Conclusions

CHAPTER 4

4.1 Introduction to Study 3

4.1.1 Perception's Influence on Production

4.1.2 Production's Influence on Perception

4.1.3 The Current Study

4.2 Method

4.2.1 Analysis

4.2.1.1 Base production models

4.2.1.2 Base perception models

4.2.1.3 Models of perception's influence on production

4.2.1.4 Models of production's influence on perception

4.3 Results

4.3.1 Perception's influence on production

4.3.2 Production's influence on perception

4.3.2.1 Looks to the target

4.3.2.2 Looks to the competitor

4.4 General Discussion

4.5 Conclusions

CHAPTER 5

5.1 The Shape of Phonetic Variation Depends on Linguistic Experience and Word Class

5.2 Phonetic Variation Benefits the Listener, Regardless of Linguistic Experience

5.3 Production and Perception of Variation is Related within Individuals

5.4 Conclusions

REFERENCES

APPENDICES

Appendix A. Stimuli

Appendix B. Measurement Criteria for Acoustic Analyses in Study 1

Appendix C. Model Output for Control Factors in Study 1

Appendix D. Model Output for Looks to the Target in the Late Window in Experiment 1 of Study 2

Appendix E. Model Output for Looks to the Competitor in the Late Window in Experiment 1 of Study 2

Appendix F. Model Output for Looks to the Target in the Late Window in Experiment 2 of Study 2

Appendix G. Model Output for Looks to the Competitor in the Late Window in Experiment 2 of Study 2

Appendix H. Output of Production Model in Study 3

Appendix I. Output of Perception Model for Looks to the Target in Study 3

Appendix J. Output of Perception Model for Looks to the Competitor in Study 3

LIST OF FIGURES

Figure 2.1. Example array from experiment interface. …..……...42

Figure 2.2. Mean response time (ms) across groups and discourse conditions. Error bars show standard error. …….…….49

Figure 2.3. Mean duration of nouns (ms) across groups and discourse conditions. Error bars show standard error. …..……51

Figure 2.4. Mean duration of determiner vowels (ms) across groups and discourse conditions. Error bars show standard error. …..……..55

Figure 3.1. Example visual display from experiment. ……..…..…..…..…..……93

Figure 3.2. Proportion of looks to the target in the late window across groups (horizontal), discourse conditions (vertical), and reduction conditions (shape) in Experiment 1. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-new). Error bars show standard error. …..……...102

Figure 3.3. Proportion of looks to the competitor in the late window across groups (horizontal), discourse conditions (vertical), and reduction conditions (shape) in Experiment 1. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., unreduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., unreduced and discourse-new). Error bars show standard error. ……..……...110

Figure 3.4. Proportion of looks to the target in the late window across discourse conditions and reduction conditions in Experiment 2. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-new). Error bars show standard error. ..……120

Figure 3.5. Proportion of looks to the competitor (empirical logit transformed) in the late window across discourse conditions and reduction conditions in Experiment 2. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., unreduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., unreduced and discourse-new). Error bars show standard error. …..……..121

Figure 4.1. Reduction effect size (discourse-new – discourse-given word durations) by BLUPs from reduction effect at cubic term for looks to the competitor in the perception experiment. Regression line shows simple linear regression with 95% confidence interval……144

Figure 4.2. Proportion of looks to the target separated by group (horizontal), discourse conditions (vertical), and reduction conditions (shape). Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and discourse-

given) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-new). Error bars show standard error……..146

Figure 4.3. Proportion of looks to the competitor separated by group (horizontal), discourse conditions (vertical), and reduction conditions (shape). Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and discourse-new) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-given). Error bars show standard error……..148

LIST OF TABLES

Table 2.1. Language background information. Mean (standard deviation) ……..……..40

Table 2.2. Summary of regression results (direction of effects) for control factors in the main analyses. Rows correspond to control variables considered for inclusion in models for each dependent measure (in columns). Direction of effects only shown for significant and marginal effects. Italicized text indicates a marginal effect. Grey indicates the variable was not included in the model ……..……...48

Table 2.3. Summary of regression results (direction of effects) for factors of interest in the main analyses. Rows correspond to variables considered for inclusion in models for each dependent measure (in columns). Direction of effects only shown for significant and marginal effects. Italicized text indicates a marginal effect ……..……..48

Table 3.1. Language background information. Mean (standard deviation) ……..……..85

Table 3.2. Mean durations of target noun and determiner in each condition (ms). Standard deviation in parentheses ……..……...…89

Table 3.3. Mean durations of words and pauses in preamble (ms). Standard deviation in parentheses ……..……..89

Table 3.4. Hypotheses and predictions for eye movement analyses ……..……98

Table 3.5. Mean recognition error rates for L2 listeners across conditions and phases of the trial in Experiment 1. Standard error in parentheses ……..……...99

Table 3.6. Mean proportion of looks to the target in the early time window (0-199 ms) across discourse conditions, reduction conditions, and groups in Experiment 1. Standard error in parentheses ……..……..……101

Table 3.7. Mean proportion of looks to the competitor in the early time window (0-199 ms) across discourse conditions, reduction conditions, and groups in Experiment 1. Standard error in parentheses ……..……..101

Table 3.8. Mean proportion of looks to the target in the early time window (0-199 ms) across discourse conditions, reduction conditions, and groups in Experiment 2. Standard error in parentheses ……...…….119

CHAPTER 1

1.1 Introduction

The opposing forces of variation and stability are central to theories of linguistic processing. Despite possessing stable linguistic representations (that constitute the knowledge necessary to use language), speakers produce tremendous amounts of variation in, for example, the phonetic properties of words. Critically, this variation is not random. Many intra-speaker and inter-speaker factors shape this variation in a meaningful way, imposing limits on the level of stability in the linguistic system. Understanding such variation therefore provides key insights into the mechanisms that shape language structure.

Early speech research focused heavily on understanding how listeners deal with variability in production; given our stable representations, how do listeners discard this variation to perceive invariant linguistic units that map to abstract representations? Decades of research has shown us that listeners do not need to disregard this variation. Instead, variation in the linguistic signal provides useful information to the listener that can facilitate, rather than impede, speech perception.

This dissertation focuses on phonetic variation in speech that is associated with the probabilistic distribution of words. Specifically, it focuses on discourse-dependent probability, which is the probability associated with a word due to its discourse status (i.e., discourse-given words have high probability, while discourse-new words have low probability). When a word has high discourse-dependent probability, it tends to be phonetically reduced (e.g., has shorter duration overall, shorter vowel durations, more centralized vowels) compared to the same word when it has low discourse-dependent probability (e.g., Fowler & Housum, 1987; Kahn & Arnold, 2015). Phonetic variation can also stem from differences across speakers, such as the level of

linguistic experience a speaker possesses (e.g., Baker, Baese-Berk, Bonnasse-Gahot, Kim, Van Engen, & Bradlow, 2011). In this dissertation, we consider the interplay of these intra-speaker and inter-speaker forces on the production of phonetic variation. Specifically, we ask how linguistic experience impacts the processing of probabilistic information during speech production.

During speech perception, phonetic variation, including the variation resulting from the processing of probabilistic information, can facilitate word recognition. This variation, coupled with the knowledge that it resulted from the processing of a high vs. low probability word, is also a type of probabilistic information. This probabilistic information serves as a predictive cue to listeners, allowing them to predict the identity of a word when presented with ambiguous input (Dahan, Tanenhaus, & Chambers, 2002). As in speech production, linguistic experience may also impact processing of probabilistic information during speech perception. That is, L1 and L2 listeners may differ in how predictive processing is engaged during speech perception (Kaan, 2014).

L2 listeners' ability to engage predictive processing may be dependent on their ability to produce the cues that drive prediction (e.g., Hopp, 2013). That is, individual differences in linguistic experience within the L2 group may impact the processing of probabilistic information in similar ways across speech production and speech perception. Researchers have long endeavored to understand the relationship between production and perception, and recent theories argue for a tight coupling between modalities, especially in the realm of prediction (e.g., Dell & Chang, 2014; Pickering & Garrod, 2013). These theories argue that listeners are able to make predictions in perception by engaging their own production system. Furthermore, given that perception has also been shown to influence production (e.g., Bradlow, Pisoni, Akahane-

Yamada, & Tohkura, 1997), the production of predictive cues (e.g., discourse-dependent reduction) is likely related to one's prediction ability. In considering the role of linguistic experience in the processing of probabilistic information, we can observe substantial variation across L2 speakers and listeners. These individual differences allow for an investigation of the relationship between production and perception.

We devote the remainder of this chapter to introducing three studies that comprise this dissertation. In these studies, we consider the role of linguistic experience in how probabilistic information influences processing during speech production (Study 1), speech perception (Study 2), and across modalities (Study 3). The remainder of this chapter introduces existing evidence motivating each study. We outline the logic behind each study and provide a preview of their results. Finally, we consider the insights this dissertation provides into the role of phonetic variation in linguistic processing.

1.2 Study 1: Linguistic Experience Influences the Processing of Probabilistic Information during Function but not Content Word Production

Processing during both content and function word production is sensitive to probabilistic information associated with the word itself, with surrounding words, or both (e.g., Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). The influence of discourse-dependent probability has been well-attested in studies of conversational speech (e.g., Aylett & Turk, 2004; Bell et al., 2009) and laboratory studies that establish a discourse in the context of the experiment (e.g., Baker & Bradlow, 2009; Lam & Watson, 2010). Systematic phonetic variation, particularly in the form of phonetic reduction on word durations, has been associated with words with high vs. low discourse-dependent probability.

Event description tasks (also called referential communication tasks) have been widely used to investigate the influence of discourse-dependent probability on phonetic reduction (e.g., Lam & Watson, 2010; Kahn & Arnold, 2012, 2015). In this task, which we utilized in Study 1, speakers see pictures undergoing a series of actions and must describe these actions in complete sentences (e.g., *The candy rotates*). Discourses are simulated within each trial, which include multiple actions on a set of pictures that remain constant during a trial. Pictures that undergo two distinct actions have high discourse-dependent probability at the second mention, and studies have reliably observed reduction in the duration of both nouns and determiners in descriptions with high vs. low probability pictures (Kahn & Arnold, 2012, 2015). However, Bell et al. (2009) failed to find significant reduction of repeated function words in a corpus of spontaneous speech.

Study 1 of this dissertation considers how linguistic experience influences the processing of probabilistic information (i.e., discourse-dependent probability) during content and function word production. We aim to test two hypotheses. The overarching hypothesis argues that L2 speakers experience deficits (compared to L1 speakers) in the processing of probabilistic information during production. Previous studies have observed differences between L1 and L2 speakers in the processing of other types of probabilistic information in picture naming tasks (e.g., lexical frequency; Gollan, Slattery, Goldenberg, Van Assche, Duyck, & Rayner, 2011), while one study found that discourse-dependent probability influences L1 and L2 productions in similar ways for read speech (Baker et al., 2011). In Study 1, we utilize a more ecologically valid speech production task (the event description task outlined above) to investigate whether differences between L1 and L2 speakers may emerge when L2 speakers must generate linguistic messages on the fly (as found for accented speech by Gustafson, Engstler, & Goldrick, 2013).

Our second hypothesis pertains to the mechanisms underlying the reduction of function words preceding content words with high discourse-dependent probability. Unlike previous studies that have argued function word reduction should be attributed to facilitated production due to priming of the recently repeated noun (Kahn & Arnold, 2012), we hypothesize that this reduction occurs because the function word inherits the probability of the following content word. To investigate this hypothesis, we leverage the interplay between linguistic experience and the processing of probabilistic information. For probability inheritance to occur, we assume that articulatory planning of the determiner is still underway during planning of the noun (Pluymaekers, Ernestus, & Baayen, 2005), allowing the discourse-dependent probability associated with the noun to influence articulation of the determiner. However, the increased demand of L2 processing, combined with the high demand of the event description task, likely limits the scope of planning (Ferreira & Swets, 2002), which blocks probability inheritance. This hypothesis predicts that L1 and L2 speakers will produce different levels of function word reduction in high vs. low probability conditions.

We observed that linguistic experience had distinct influences on the processing of probabilistic information during content vs. function word production. L2 speakers produced significantly longer word durations overall compared to L1 speakers, replicating a series of previous studies (e.g., Baker et al., 2011). Despite this overall difference between groups, both L1 and L2 speakers reduced high vs. low probability content words to a similar degree. Similar effects were shown for response times (an index of planning). This indicates a strong influence of probabilistic information on L2 processing even during a demanding production task. Critically, a different pattern of results was found for production of function words. L2 function word durations did not differ substantially from "standard" L1 durations (i.e., function words

produced in low probability contexts), suggesting sufficient mastery of English prosody for common function words (see also, Baker et al., 2011). However, only L1 speakers significantly reduced their determiners when the following noun had high vs. low discourse-dependent probability, a result predicted by the probability inheritance hypothesis.

1.3 Study 2: Linguistic Experience Impacts the Time-Course of Prediction during Speech Perception

The coupling of discourse-dependent probability and reduction constitutes probabilistic information that could be useful to listeners during speech perception. As discussed above, L1 speech provides this probabilistic information to listeners for both content and function words. Previous work has shown that when this information is associated with content words, it serves as a predictive cue to listeners. When receiving incoming ambiguous input about a content word, the congruent coupling of discourse-dependent probability and reduction facilitates the recognition of that word (Arnold, 2008; Dahan et al., 2002; Isaacs & Watson, 2010). This leaves the question of whether probabilistic information associated with the previous function word allows listeners to predict the upcoming content word. This sort of predictive processing has been observed with L1 listeners for disfluent determiners (Arnold, Tanenhaus, Altmann, & Fagnano, 2004), gender-marked determiners (Dahan, Swingley, Tanenhaus, & Magnuson, 2000), and determiners with coarticulatory cues (Salverda, Kleinschmidt, & Tanenhaus, 2014). For example, when listeners heard disfluent determiners, they predicted that the upcoming noun was discourse-new vs. discourse-given (because disfluency occurs more often for discourse-new vs. discourse-given referents; Arnold & Tanenhaus, 2007).

Sufficient linguistic experience is critical for the engagement of prediction during speech perception for a few reasons. Extensive exposure to the language is necessary for listeners to be

able to recognize the information that can serve as predictive cues. In many cases, mere exposure is not enough; listeners must have mastered some structure of the language in order to make predictions based on that structure. For example, Hopp (2013) found that L2 German listeners (L1-English) were only able to make L1-like predictions based on the gender marking on determiners (e.g., predict a feminine noun upon hearing a feminine determiner) if they could consistently produce gender marked determiners. Even when L2 listeners successfully engage predictive processing, it may differ qualitatively from the predictions made by L1 listeners. Dijkgraff, Hartsuiker, and Duyck (2016) found that L2 English listeners (L1-Dutch) were slower than L1 listeners to predict nouns based on semantic cues from the verb (e.g., *reads* and *letter* vs. *steals* and *letter*). This set of results shows how differences in L1 and L2 experience and/or difficulties in L2 speech perception more generally (Kaan, 2014) create difficulties in L2 predictive processing.

The L2 listeners in Study 2 were the same individuals who participated in Study 1. Therefore, we have some information about the experience these listeners have with discourse-dependent probabilistic information in terms of their production behavior. Because these individuals produced L1-like levels of discourse-dependent probabilistic reduction (see also Baker et al., 2011), they may also be able to use this probabilistic information to make predictions during perception (as was the case for gender marked determiners for listeners in Hopp, 2013). However, if the L2 listeners engage predictive processing, it may differ qualitatively from L1 predictive processing (e.g., Dijkgraff et al., 2016; Kaan, 2014). For probabilistic information associated with determiners, though, it seems unlikely that L2 listeners will be able to use this information as a predictive cue, as they did not produce this type of variation in Study 1.

Study 2 of this dissertation investigates whether L1 and L2 listeners show similar use of probabilistic information as a predictive cue during speech perception. We consider whether these groups of listeners make predictions based on probabilistic information associated with both determiners and the following nouns. To test these possibilities, we used a visual world eye-tracking experiment, in which listeners followed a series of instructions to move objects in the visual display (e.g., *Put the candy below the square… Now put the candle above the diamond*). Trials either included a target with low discourse-dependent probability (i.e., discourse-new, *candle* in the preceding example) or with high discourse-dependent probability (i.e., discourse-given, if *candle* appeared in both instructions in the preceding example). Targets were either reduced or unreduced, creating congruent or incongruent coupling of discourse-dependent probability and reduction in each trial, which provided listeners with potential predictive cues. Evidence that listeners engage predictive processing comes from sensitivity to (in)congruency – e.g., overall higher proportion of looks to reduced vs. unreduced targets with high discourse-dependent probability; a shallow drop off in looks to the cohort competitor (e.g., *candy* in the above example) when the high probability target was unreduced vs. reduced.

The results of Study 2 revealed that probabilistic information associated with the target noun allowed both L1 and L2 listeners to make predictions about the identity of the noun in the face of ambiguous input. While both groups engaged predictive processing, L1 and L2 listeners differed in how these predictions unfolded over time. L2 listeners made strong predictions when the probabilistic information was consistent with the competitor rather than the target, and found it difficult to recover when bottom-up input confirmed the identity of the target (proving their predictions had been incorrect). These failed predictions sometimes led to recognition errors for L2 listeners, while L1 listeners never selected the competitor over the target. In sum, L2 listeners

can engage predictive processing, but there are qualitative differences in the predictions made by L1 and L2 listeners. Neither group of listeners used probabilistic information associated with the determiner to predict the identity of the upcoming noun; this may reflect issues with the acoustic properties of our experimental stimuli.

1.4 Study 3: Similarities in How Probabilistic Processing Influences Production and Perception: Cross-Task Transfer

Recent theories of prediction argue that predictions are driven, at least in part, by production processes (Dell & Chang, 2014; Pickering & Garrod, 2013). For production to drive prediction, there must be a tight relationship between speech production and speech perception that allows mutual influence across modalities. Past research has considered these influences, and found that perception influences production, especially via perceptual learning, and that production influences perception, especially via prediction. For example, Kittredge and Dell (2016) demonstrated that transfer from perception to production occurred during training on a new phonotactic constraint (e.g., /f/ can only occur in coda position), but only when the perception task involved some sort of production processes (e.g., inner speech). Similarly, the engagement of predictive processing is thought to involve production. Consistent with this idea, Hopp (2013; reviewed above) found that a certain level of production ability was necessary in order to engage predictive processing.

Study 3 leverages individual differences in reduction and prediction ability across the L2 participants in Studies 1 and 2 in order to investigate links between speech production and speech perception. We are particularly interested in the relationship between modalities in terms of how probabilistic information influences processing. In a re-analysis of Study 1, we ask whether indices of listeners' sensitivities to reduction in perception relate to reduction in

production. A re-analysis of Study 2 complements this by examining whether indices of the influence of probabilistic information on production – the reduction of discourse-given vs. discourse-new nouns – relate to listener's sensitivity to discourse-dependent probabilistic reduction in perception.

The results of Study 3 reveal that listeners who show strong prediction effects in perception also produce large degrees of reduction in production. Furthermore, speakers who produced large degrees of reduction for nouns with high vs. low discourse-dependent probability were the individuals who exhibited strong prediction effects during speech perception. Those who produced little or no reduction showed weaker or no prediction effects. Together, these results support strong links between production and perception in terms of how probabilistic information influences processing.

1.5 Conclusions

The three studies that comprise this dissertation add to our understanding of the factors that condition phonetic variation. By manipulating the level of linguistic experience of our speakers and listeners, we accomplished much more than a mere investigation of how L1 and L2 speech processing differ. In Study 1, we found that L1 but not L2 speakers reduce the duration of determiners preceding nouns with high vs. low discourse-dependent probability, which sheds light on the mechanisms underlying the influence of probabilistic information on function word reduction. In Study 2, we demonstrated that L2 listeners use discourse-dependent probabilistic reduction to make predictions during speech perception, but that they cope poorly with prediction error compared to L1 listeners. This difference raises questions about the mechanisms that drive prediction, fueling future research. Finally, in Study 3, we discovered transfer between the ability to produce predictive cues and the ability to use those cues to make predictions, which

refines our understanding of the links between production and perception. None of these new insights and questions would have come about without comparing the behavior of individuals with different levels of linguistic experience. This dissertation showcases the benefits of bilingual research and paints a detailed picture of how probabilistic information influences speech processing both within and across modalities.

CHAPTER 2

2.1 Introduction to Study 1

Many variables shape variation in the acoustic-phonetic properties of words, including inter- and intra-speaker factors. Striking differences in speech production behaviors, including the acoustic-phonetic characteristics of productions, can be observed when comparing native, first language (L1) speakers of a language to non-native, second language (L2) speakers. For example, L2 speakers produce overall longer word durations than L1 speakers (Munro & Derwing, 1995). However, beyond this inter-speaker variation, variation in the duration of speech can also be observed within an individual during the production of both function and content words. For example, the probability of a word occurring in a discourse impacts planning and articulation. Existing research has shown that content words that have high probability within a particular discourse are typically planned more quickly (e.g., Kahn & Arnold, 2012) and are reduced in duration (e.g., Fowler & Housum, 1987) relative to low probability words. Interestingly, the influence of probabilistic information on processing may differ across word classes. The robust effects of probability within a discourse observed for content words are inconsistently found for function words (e.g., Bell, Brenier, Gregory, Girand, & Jurafsky (2009) find no effect, where Kahn & Arnold (2012, 2015) find significant reduction of high probability function words).

In the current work, we examine how the processing of such probabilistic information across word classes is influenced by linguistic experience. Because L2 speakers have less experience with a language than L1 speakers, L2 speech processing is more demanding. This increased demand is thought to contribute to overall differences in production behavior across groups (e.g., longer L2 content word durations; Baker, Baese-Berk, Bonnasse-Gahot, Kim, Van

Engen, & Bradlow, 2011). The current study examines how these increased processing demands, accompanied by overall differences in linguistic experience (e.g., lack of familiarity with the prosodic structure of the L2), influence the manner in which speakers are able to utilize probabilistic information during planning and articulation for both content and function words. We begin by discussing how probabilistic information influences processing, and how it may impact content and function word processing differently. Next we turn to existing work on the influence of probabilistic information on processing during content (specifically, noun) and function (specifically, determiner) word production by both L1 and L2 speakers. Finally, we consider how differences among L1 and L2 speakers may shed light on the mechanisms underlying the processing of probabilistic information.

2.1.1 The Processing of Probabilistic Information in Speech Production

The probability of some linguistic unit can have a demonstrable influence on the processing taking place at some level of the speech production system. For example, verbs have biases toward structures with which they are typically associated (i.e., high probability structures). When a speaker selects a structure for the verb phrase containing this verb, probabilistic information comes into play. Gahl and Garnsey (2004) found that the duration of verbs and their arguments were significantly shorter when a high vs. low probability syntactic structure was selected.

Variation in probability at many levels of linguistic structure can influence phonetic outcomes during speech production (e.g., word durations). The current study focuses on the inverse relationship between the probabilities of individual words and their phonetic prominence. The distribution of a word (i.e., its occurrence within continuous speech) can be modeled by considering its probability with respect to a range of factors or various levels of linguistic

structure. For example, the frequency of some word is conditioned on its distribution within a particular language and the discourse-dependent probability of a word is conditioned on its distribution within some particular discourse. Acoustic-phonetic reduction of some kind, whether it be in the duration of the stressed vowel of the word, the spectral qualities of the vowel, or the duration of the word itself, has been associated with high vs. low probability words (e.g., Aylett & Turk, 2004; Baker & Bradlow, 2009; Bell et al., 2009; Fowler & Housum, 1987; Gahl, 2008; Gahl & Garnsey, 2006; Jurafsky, Bell, Gregory, & Raymond, 2001; Kahn & Arnold, 2015; Lam & Watson, 2010; Liberman, 1963; Munson & Solomon, 2004; Pate & Goldwater, 2015; Scarborough, 2010). Reduction in the time required to plan a word or utterance (indexed by response time, or RT) is also impacted by word-specific probability (e.g., Kahn & Arnold, 2012, 2015). In the current study, we examine the association of both measures – planning time and acoustic-phonetic reduction – to discourse-dependent probability.

2.1.2 Processing of Content vs. Function Words

Words clearly vary in the richness of their semantic and syntactic properties. A fundamental contrast that has long been noted in linguistic theories is between semantically rich, but syntactically weak, content words and syntactically rich, but semantically impoverished, function words (although this dichotomy is likely an oversimplification; c.f. Altmann, Pierrehumbert, & Motter, 2009). Theories of word production typically consider these word classes separately (Levelt, Roelofs, & Meyer, 1999). Some theories go as far as to say that determiners are not lexical entities, and thus do not undergo lexical selection processes (as content words do; Garrett, 1975). According to such theories, function words are retrieved as part of some syntactic structure, or frame. For example, during production of an English noun phrase, the determiner frame is retrieved with its head, already filled; in contrast, the head of the

noun phrase frame is empty, which must then be filled by lexical selection processes. In contrast, more recent theories argue that function words, such as determiners, have lexical representations and undergo similar selection processes as for content words (Bürki, Laganaro, & Alario, 2014; Janssen, Schiller, & Alario, 2014; Jescheniak, Schriefers, & Lemhöfer, 2014). Under either type of theory, the selection of a determiner depends upon (or, occurs after) selection of the noun phrase's head noun (Bock & Levelt, 1994), as the specific determiner to be selected (e.g., *a* vs. *an* in English, or *le* vs. *la* vs. *l'* in French) depends upon the form of the following noun (Caramazza, Miozzo, Costa, Schiller, & Alario, 2001).

While psycholinguistic theories of production typically assume that function and content words share common phonological encoding and phonetic implementation processes following selection (Lapointe & Dell, 1989), evidence from L1 speech suggests that function and content words differ in their phonological form and prosodic implementation. Unlike content words, which are typically stressed, (monosyllabic) function words tend to be unstressed unless produced in isolation (Selkirk, 1996) or when following a disfluency (Bell, Jurafsky, Fosler-Lussier, Girand, Gregory, & Gildea, 2003). Due to their "weak" phonological form, function words cannot form a foot (a prosodic structure made up of at least one strong, accented syllable) or a prosodic word (a prosodic structure consisting of at least one foot). Therefore, without a prosodic word of their own, function words cliticize to an adjacent content word to form a single prosodic word (Selkirk, 1996).

While L1 and L2 speech production are assumed to rely on the same types of selection processes for all word classes, L2-specific factors influence the dynamics of these processes. For content word production, L2 speech is often found to have overall longer word durations than L1 speech (Baker et al., 2011; Guion, Flege, Liu, & Yeni-Komshian, 2000; Munro & Derwing,

1995). This across-the-board slowing can be attributed to two non-mutually exclusive sources: cross-language interference or impoverished linguistic knowledge.

According to the cross-language interference account, both L1 and L2 lexical representations are automatically activated in parallel during L2 speech production (e.g., Colomé, 2001; Costa, Caramazza, & Sebastian-Galles, 2000; Colomé & Miozzo, 2010; Hermans, Bongaerts, De Bot, & Schreuder, 1998). Delays in both speech planning (e.g., Hermans et al., 1998) and articulation (Sadat, Martin, Alario, & Costa, 2012) are consequences of this processing difficulty attributed to cross-language interference.

Additional processing difficulty could arise because L2 speakers necessarily have less experience with the language compared to L1 speakers. One proposal, the frequency lag hypothesis (Gollan, Slattery, Goldenberg, Van Assche, Duyck, & Rayner, 2011), argues that L2 speakers have accrued lower frequency counts for all words than L1 speakers due to their lower level of exposure to the language. Slower speech articulations for L1 compared to L2 speakers could, therefore, be due to the L2 frequency lag, given that lower frequency words have longer word durations (e.g., Bell et al., 2009). That is, any given word produced by an L2 speaker should be longer than if produced by an L1 speaker simply because it has lower frequency.

Duration differences between L1 and L2 speech also manifest in function word production. A number of studies have found that L2 speakers fail to reduce function words to the same degree as L1 speakers (Aoyama & Guion, 2007; Baker et al., 2011). This is true even for speakers who would reduce function words in their own L1 (e.g., Mandarin; Shi, Morgan, & Allopenna, 1998). This difference has been attributed to a lack of mastery of the English prosodic system (Baker et al., 2011), where L2 speakers fail to encliticize function words to adjacent prosodic words (Selkirk, 1996) and instead give them independent status as prosodic

words. This difficulty may reflect broader differences between English and Mandarin prosodic structure. English, unlike Mandarin has alternating stressed and unstressed syllables (enclitization is one type of structure that leads to such alternations). Therefore, the prosodic structure of English may be particularly challenging for Mandarin speakers to master. This may be especially difficulty for the definite determiner, as Mandarin lacks this type of function word.

However, it is worth noting that the function word differences found by Baker et al. (2011) only manifested for the lower frequency function words; for example, the L2 speakers reduced *the* (the third highest frequency (function) word in English) to the same degree as L1 speakers. This suggests that L2 speakers accrue enough experience with high frequency words to reduce them to an L1-like degree (consistent with the frequency lag hypothesis). With enough experience with a particular word, L2 speakers may then be able to overcome general issues with English prosody.

Altogether, across L1 and L2 speech, we can see marked differences in the functional properties (i.e., semantically vs. syntactically focused) and prosodic implementations of content and function words. As discussed in more detail below, studies also observe distinct influences of probabilistic information across these word classes. In the following sections, we discuss these diverging results and consider possible differences that might be present across L1 and L2 speakers stemming from general differences in speech production behavior across groups, as discussed above.

2.1.3 Effects of Variation in Content Word Probabilities

2.1.3.1 L1 speech. Influences of probabilistic information on content word production have been well-established in the literature for L1 speech production. Bell et al. (2009) analyzed a corpus of spontaneous, conversational speech and found that various types of probabilistic

information influenced word duration. Words with higher lexical frequency and higher preceding/following conditional probability (i.e., the probability of a word given the previous or following word) were reduced relative to lower frequency/probability words. Critically, controlling for these factors, content words with high discourse-dependent probability were significantly reduced. Robust effects of discourse-dependent reduction have also been observed in experimental tasks that simulate spontaneous conversation and/or create a discourse within the context of the experiment, such as interactive map tasks (Aylett & Turk, 2004; Bard, Anderson, Sotillo, Aylett, Doherty-Sneddon, & Newlands, 2000; Meagher & Fowler, 2014), referential communication tasks (Jacobs, Yiu, Watson, & Dell, 2015; Kahn & Arnold, 2012, 2015; Lam & Watson, 2010, 2014), and paragraph reading (Baker & Bradlow, 2009; Baker et al., 2011).

For example, in event description tasks (also called referential communication tasks), target speakers perform an augmented picture naming paradigm in which they describe a series of events occurring with a set of pictures presented on the screen (e.g., *The candle rotates* or *The candy goes on the red*). Critical trials include two events with the same picture, thus, at the second production of the picture name, the noun has high discourse-dependent probability. A number of studies utilizing variants of this paradigm have found robust reduction effects both in RTs and word durations stemming from the processing of discourse-dependent probabilistic information. Using this event description paradigm, Lam and Watson (2010, 2014) established that speakers reduced the production of nouns independently of whether the noun otherwise had high probability in a trial, and that the probability of the lexical item (i.e., the noun) is critical for reduction rather than probability of the particular entity to which the noun refers. Subsequent studies with this paradigm have considered how discourse-dependent probability influences RTs, in addition to target noun durations. Collectively, these studies established that a word can attain

high probability by a number of means, including introduction of the word to the discourse either linguistically (heard or produced) or non-linguistically (i.e., visually; Kahn & Arnold, 2012, 2015; Jacobs et al., 2015). Under these conditions, all studies observed reduced RTs and noun durations when the target was discourse-given vs. discourse-new. Similar effects with content word reduction have been observed in tasks with paragraph reading, which also establish a discourse in the context of the experiment (Baker and Bradlow, 2009; Baker et al., 2011).

Note that many of these studies have suggested that reduction observed in this paradigm should not be attributed to explicit computation of discourse-dependent probabilities, but rather to facilitation during planning and execution that occurs due to priming from the recently repeated nouns (Arnold & Watson, 2012; Fraundorf, Watson, & Benjamin, 2014; Kahn & Arnold, 2012; Lam & Watson, 2010, 2014; c.f., Fowler, 1988). We return to this issue below.

2.1.3.2 L2 speech. Cross-language interference and frequency lag, the two possible sources of L2 slowing during content word production, could also lead to deficits in how probabilistic information influences L2 processing. As argued above, high levels of interference from the L1 increase the demand of L2 vs. L1 speech production processing. With this overall increased demand, L2 speakers may not have the resources available to track probabilities in an L1-like manner. Similarly, less experience overall with an L2 means less exposure to probability distributions in the L2. The use of different types of probabilistic information in production should yield different patterns of performance in L1 and L2 speakers. In fact, the exposure-driven frequency lag hypothesis is motivated by empirical findings that L2 speakers exhibit probabilistic effects differing in magnitude than effects from L1 speakers. For example, variation in lexical frequency (a type of word-specific probability) has a larger effect on RTs for L2 compared to L1 speakers, both in speech production and perception (e.g., Diependaele,

Lemhöfer, & Brysbaert, 2013; Gollan, Montoya, Cera, & Sandoval, 2008; Gollan et al., 2011; Hernández, Costa, & Arnon, 2016), which indicates differences in how probabilistic processes influence L1 vs. L2 speech.

However, such differences have not always been found, particularly in phonetic measures. Differences in frequency's influence on noun durations have not been reliably observed (Baker et al., 2011; Sadat et al., 2012). For discourse-dependent reduction in word durations, Baker et al. (2011) also did not find reliable differences between groups; L2 speakers reduced discourse-given content words to the same degree as L1 speakers in read speech. This result suggests that although cross-language interference or low levels of experience may impact overall word durations and frequency effects on planning, the influence of discourse-dependent probabilities on processing remains unimpaired during L2 speech production. Therefore, it seems to be the case that L2 speakers also track the discourse status of words and reduce accordingly (for further discussion of reduction of nouns in L2 speech, see Lam & Marian, 2015).

Such a null result is consistent with the view that reduction reflects priming (e.g., Lam & Watson, 2014). Since this mechanism is shared across L1 and L2 speech, both types of speakers should exhibit comparably levels of reduction. However, read speech is considerably less demanding than typical production situations. In fact, previous findings suggest that L2 effects on word durations are mitigated in less demanding production tasks, such as single word repetition compared to picture naming (Gustafson, Engstler, & Goldrick, 2013). Therefore, the seemingly unimpaired processing of probabilistic information by L2 speakers in Baker et al. (2011) could be attributed to the paragraph reading task utilized in that study.

2.1.3.3 Study goal 1: L1 vs. L2 differences in the effects of content word probabilities in a more demanding task. Given these differences across studies of probabilistic reduction,

one untested possibility is that deficits in how probabilistic information influences processing during L1 vs. L2 production can be observed when L2 speakers engage in a more demanding task that requires them to generate full sentences without a written prompt (e.g., an event description task; Lam & Watson, 2010). The first goal of the current study is to test this hypothesis. The increased processing demands of such a task could impede the ability of L2 speakers to track the probability of words within a discourse, which could lead to differences between L1 and L2 speakers in the magnitude of discourse-dependent probabilistic reduction produced. Alternatively, the increased demand could pose little difficulty for discourse-dependent probabilistic reduction on the noun itself, but could impact probabilistic reduction of other types of words within the sentence, namely, determiners.

2.1.4 Effects of Variation in Function Word Probabilities

2.1.4.1 L1 speech. In addition to content word reduction, Bell et al. (2009) also considered the influence of different sorts of word-specific probabilities on function word production. Notably, unlike for content words, they found that function words did not undergo discourse-dependent reduction. However, function word reduction is strongly predicted by multi-word probability (i.e., probability conditioned on either the previous or following word; Bell et al., 2003; Bell et al., 2009; Jurafsky et al., 2001). For example, Bell et al. (2003) found a strong relationship between the duration of determiners and the conditional probability of the determiner given the following word. Therefore, it could be the case that discourse-dependent probabilistic reduction only occurs on function words when they modify nouns with high discourse-dependent probability. We refer to this as the *probability inheritance hypothesis*. While not tested by Bell et al. (2009), other studies have reported effects compatible with this hypothesis (Kahn & Arnold, 2012, 2015). In the same event description tasks that elicited

discourse-dependent reduction on nouns, Kahn and Arnold have reported significant reduction of function words in discourse-given vs. discourse-new trials.

As with reduction of content words, such effects could be attributed specifically to processing of probabilistic structure. If production takes into account the probability of the determiner conditioned on the following noun, the determiner could 'inherit' the high discourse-dependent probability of the noun. Alternatively, as with the reduction of content words, the reduction of determiner durations could be attributed to priming from the recently repeated nouns (e.g., Kahn and Arnold, 2012).

2.1.4.2 L2 speech. Little research has considered how probabilistic information affects L2 function word production. Schertz and Ernestus (2014) measured the duration of determiner vowels produced by Norwegian and Czech speakers of English, and asked whether variation in the frequency of the following noun influenced determiner production. Productions were taken from spontaneous conversations between L2 speakers of English (with the same L1 background) in which one speaker described the contents of a picture to the other, who was asked to replicate the picture. The results showed that the Czech, but not Norwegian, speakers produced shorter determiner vowels when followed by more vs. less frequent nouns. Therefore, like L1 English speakers (Bell et al., 2003), (at least some) L2 English speakers reduce function words based on the probability of the content words that they modify.

While Schertz and Ernestus (2014) found that probabilistic information influences L2 processing of function words, leading to phonetic reduction, the question remains as to whether L2 speakers differ from L1 speakers in the degree of reduction that they produce. This sort of direct comparison is needed to establish whether L2 speakers show deficits in how probabilistic information influences function word production.

2.1.4.3 Study Goal 2: L1 vs. L2 differences in the effects of function word probabilities in a more demanding task.

If differences are observed across L1 and L2 speech for content words, L2 function word production can help address a second goal of this study. Under the probability inheritance hypothesis, we assume that the mechanism that allows the discourse-dependent probability of content words to influence articulation of determiners requires an overlap in planning; articulatory planning of the determiner is still underway during planning of the noun (Pluymaekers, Ernestus, & Baayen, 2005). However, given that processing difficulty shortens the scope of planning at the lexical level (e.g., Ferreira & Swets, 2002), we hypothesize that similar limitations in scope can occur for articulatory planning due to the increased demand of L2 processing. This narrow scope for L2 speakers prevents determiners from inheriting the discourse-dependent probability of the nouns. Therefore, the probability inheritance hypothesis predicts that L1 and L2 speakers will reduce determiners different degrees in discourse-given vs. discourse new conditions. Alternatively, if reduction solely reflects priming of the following noun — a mechanism shared across L1 and L2 speech — we expect that L1 and L2 speakers will reduce determiners to similar degrees across discourse conditions.

2.1.5 The Current Study: Summary

The current study investigates (1) the presence or absence of L2 deficits in how discourse-dependent probabilistic information influences processing during a demanding speech production task, and (2) whether this influence on function word production occurs due to probability inheritance from the target noun. To address these two goals of the study, we use an event description task, in which speakers describe animations of a series of pictures (e.g., *The candle rotates*). We compare indices of planning and execution (RTs and word durations, respectively) across discourse-new (the target word is produced only once within a trial) and

discourse-given (the target word is produced twice) trials. A decrease in the magnitude of discourse-dependent reduction in RTs and/or content word durations speaks to the presence of difficulty in the processing of probabilistic information during L2 speech. If difficulties in L2 processing prevent speakers from simultaneously planning the articulation of determiners and nouns, the probability inheritance hypothesis predicts a reduction in L2 vs. L1 discourse-dependent effects specifically for function words.

2.2 Method

2.2.1 Participants

A total of 64 speakers participated in this study. Participants were divided into two groups: native speakers of American English (henceforth, the L1 group) and native speakers of Mandarin who learned English as a second language (henceforth the L2 group). Speakers comprising the L1 group (N = 25) were recruited from the linguistics participant pool at Northwestern University and received partial course credit for their participation. All participants in this group were L1 speakers of English with no history of speech impairments or color blindness. One participant was excluded due to poor equipment performance in a companion study not discussed here, leaving 24 native speakers of American English (17 female; mean age: 19, range: 18-21) for the L1 group.

Many speakers included in the L2 group were recruited from the International Summer Institute at Northwestern University, a month-long program for incoming international students that offers intensive English instruction and one-on-one tutoring prior to the start of their first academic quarter as graduate students. Other participants were recruited from the Northwestern community via flyers and a database of current and former students in Northwestern's English Language Learners Program. Each of these participants was compensated $10/hour. Two

participants were recruited via the linguistics department participant pool, and thus received partial course credit for their participation.

Thirty-nine individuals whose L1 was Mandarin participated as part of the L2 group. Eleven participants were excluded who were unable to produce at least 70% of the target items name during the experiment. This criterion was set to ensure 1) sufficient statistical power and 2) sufficient levels of English proficiency. An additional three participants were excluded due to poor equipment performance in a companion study, and one English-dominant speaker was excluded to ensure comparable linguistic experience across the participants. The remaining 24 participants (18 female; mean age: 23.2, range: 18-31) were Mandarin-dominant, L2 speakers of English who did not learn English at home (i.e., English exposure began at school).

Each participant completed a detailed language background questionnaire. In this questionnaire, participants were asked to provide information about their exposure and experience with all languages they spoke. Table 2.1 reports information provided by participants that summarizes variation in linguistic experience across groups.[1] The LexTale vocabulary test (an unspeeded lexical decision task; Lemhöfer & Broersma, 2012) was used as an objective measure of English proficiency.

[1] A series of analyses considering the influence of these proficiency variables on performance revealed no reliable effects. Given the small sample size of the groups and the substantial variation in the dependent measure we draw no strong conclusions from these null effects.

Table 2.1. Language background information. Mean (standard deviation).

	Percent Correct, LexTale vocabulary test	Age of first exposure to English (years)	Length studied English (years)	Percent time English used	Self-rated speaking ability (Perfect:10)	Self-rated listening ability (Perfect:10)
L1 group (N = 24)	95.7 (5.2)	0 (0)	18.9 (1.2)	95.7 (5.8)	9.7 (0.6)	9.8 (0.5)
L2 group (N = 24)	80 (13.4)	7.9 (5.8)	15.6 (6.8)	48.6 (27.4)	5.8 (2.4)	6.6 (2.3)

2.2.2 Materials and Design

The stimuli were a set of 48 pictures taken from the Bank of Standardized Stimuli (BOSS; Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010) and other sources as needed. The BOSS database includes a large set of full color photographs that have been normed along a number of dimensions, including name agreement, category, familiarity, and visual complexity. The names for the target pictures have high name agreement by L1-English speakers (mean = 94.6%, min = 63.2%, sd = 9.2%), as established in a separate norming study with 19 participants. The names also have above at least 40% name agreement by L2-English speakers from the same population as the participants for the current study, also established in a separate norming study with 10 participants (mean = 77.9%, sd = 20.3%).

The target items comprise 24 pairs in which the items overlap by least two phonemes (e.g., /kæn-/ in *candy* and *candle*) and have the same number of syllables. Each of these pairs was assigned three other items that served as non-target distractors. The non-targets were not semantically or phonologically related to either target in the pair. Each item in the pair differed substantially in frequency per million in the SUBTLEX-US corpus (Brysbaert & New, 2009) and covered a broad range of frequencies (from 0.2 occurrences per million and 509.4 occurrences

per million; mean = 43.4, sd = 77.7). While not the primary focus of the study, this frequency range allowed for an examination of the influence of lexical frequency on word durations. In addition to the 48 target and 72 non-target pictures, 144 filler pictures were selected from BOSS. Care was taken to avoid overlap with the target and non-target pictures, and to select images that L2 speakers were likely to be able to name successfully (e.g., *bird* but not *saxophone*). See Appendix A for the full set of stimuli, with name agreement and frequency data for each target item.

The participants' task was to describe events presented on a computer screen. At the start of a trial, an array of eight pictures appeared on the screen (see Figure 2.1). On target trials, the array included the target picture and one, two, or all three of the assigned non-target distractors. The array of pictures did not include the other target item in the pair to guard against influences of phonological overlap from phonologically related competitors on target productions (Meyer & Damian, 2007; Morsella & Miozzo, 2002). The remaining four to six pictures in each array were randomly selected from the entire set of target and filler pictures.

The experiment included 144 trials, including 96 target trials and 48 filler trials. Each target trial included three or four events (48 trials of each length). Variable trial lengths were implemented to discourage participants from adopting a prosodic strategy of utterance-final lengthening for descriptions of the third (critical) event, which could counteract the effect of discourse-dependent reduction. In each event within a trial, the pictures underwent an action: *expand*, *rotate*, *shrink*, or *fade*. Within a trial, each event involved a different, randomly selected action. For discourse-given target trials, the first and third events occurred with the same target picture. The second event occurred with a non-target picture. For discourse-new trials, all events occurred with different pictures (the target and two or three of the assigned non-targets). The

target picture in discourse-new trials always appeared in the third event, for comparison with the third event in discourse-given trials. In a discourse-given trial as illustrated in Figure 2.1, the target item (*tie*) would occur in the first event (e.g., *The tie rotates*). Next, a non-target item (*kite*) would occur in the second event (e.g., *The kite shrinks*). Finally, the target item would occur in the third event with a distinct action (e.g., *The tie fades*).

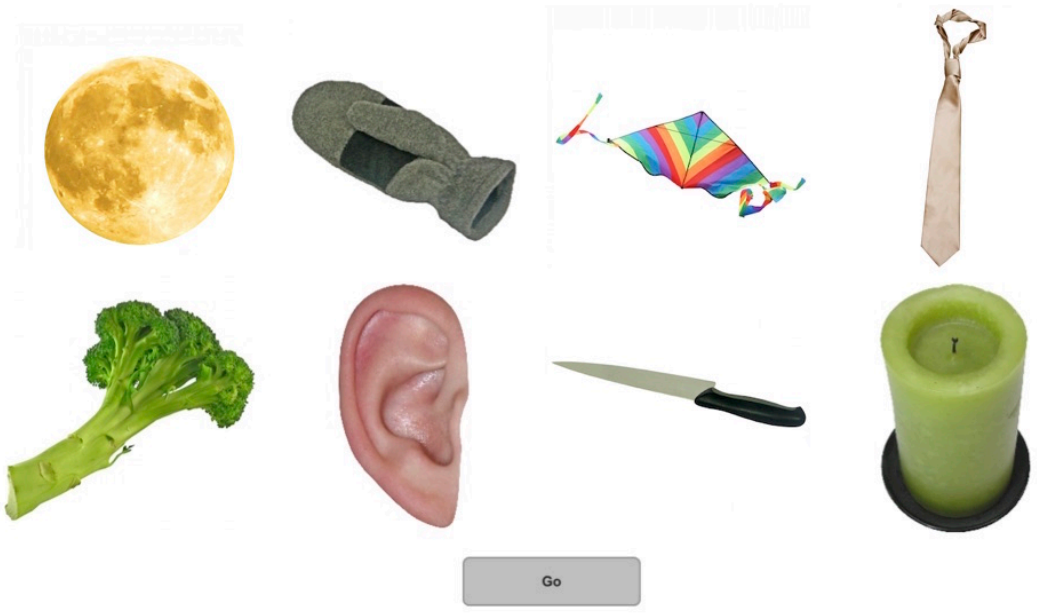

Figure 2.1. Example array from experiment interface.[2]

So that all comparisons were within item and within participant, each participant produced each target word twice over the course of the experiment (once in a discourse-given trial and once in a discourse-new trial). These conditions were blocked such that for any target word, the discourse-given and discourse-new trials appeared in different blocks (separated by breaks in the experiment). Items were also assigned to sub-blocks such that the two targets in a

[2] Images are from the Bank of Standardized Stimuli (BOSS; Brodeur et al., 2010) and are authorized for redistribution according to the Creative Commons Attribution-Share Alike 3.0 license (https://creativecommons.org/licenses/by-sa/3.0/).

pair (e.g., *tie* and *tire*) occurred in separate sub-blocks. Therefore, blocks included trials of both conditions, although never for the same target item or target pair. Finally, each block contained 12 filler trials, which ranged in length from two to four events. All filler trials were discourse-new, and the pictures were composed into sets by choosing at random from the set of candidate filler pictures. Each filler picture appeared in only one filler trial.

Eight lists were created to counterbalance the sub-block assignment of each target picture (1A, 1B, 2A, or 2B), discourse condition order (discourse-given in block 1 vs. discourse-new in block1), and trial length (3 or 4). Trials of all lengths were evenly distributed throughout the experiment, such that each sub-block (of 36 trials) contained four length-2 filler trials, four length-3 filler trials, four length-4 filler trials, 12 length-3 target trials, and 12 length-4 target trials. For each event in a trial (both target and filler), an action was chosen at random with no repetition of actions within a trial. Three versions of each list were generated with different within-block random orders, creating 24 experiment lists.[3] Each participant was randomly assigned to one of these lists, with no more than one participant from each group assigned to each list.

2.2.3 Procedure

Participants were seated in a sound-attenuated booth in front of a computer screen. They were told that they would see an array of objects and were instructed to describe the actions that occurred with those objects as soon as they recognized the action. They were familiarized with

[3] An error in the counterbalancing of one list was discovered during testing. In length 4 trials, all fourth events used the same action (*rotate*), leading to some repetition of that action within a trial. However, because the fourth event of trials were never analyzed, the general structure of the list was still acceptable for the purposes of the experiment. Therefore, data from this participant was included. The error was rectified for the other two randomized versions of this list.

the desired names for these actions (i.e., *rotate, shrink, expand,* and *fade*), and were told that each trial would include two, three, or four events. Participants were instructed to describe the event using a complete sentence, such as *The dog rotates*, and were asked to use single word names for the pictures (e.g., *kettle* but not *tea kettle*, or *ball* but not *blue ball*). The experimenter demonstrated the structure of the experiment and the desired type of description with two trials. Then, the participants completed four practice trials under the supervision of the experimenter, with any mistakes corrected. Participants were permitted to complete the practice trials an additional time upon request.

Following the practice trials, participants began the experiment, which was presented using Max/MSP software (Puckette, Zicarelli, Sussman, Clayton, Bernstein, Nevile, Place, Grosse, Dudas, Jourdan, Lee, & Schabtach, 2011). Participants controlled the initiation of events by clicking a *Go* button on the screen. Upon clicking *Go*, a grid of pictures appeared. After a two second delay, the first event occurred. When the participant finished speaking and was ready for the next event, they clicked *Go* again. The second event occurred after a 500 ms delay. Clicking *Go* the third time either elicited a third event (after a 500 ms delay), or a new grid of pictures appeared and the sequence restarted. On some trials, a fourth event occurred after clicking *Go* the fourth time (again, after a 500 ms delay). There were three breaks during the experiment, each occurring after 36 trials. Participants determined the length of the breaks. Recordings were made with a boom-mounted Shure SM81 Condenser Handheld Microphone sampling at 44,100 Hz.

2.2.4 Measurement

Speech onset latencies and determiner vowel durations and target noun durations in critical productions were manually annotated and measured using Praat (Boersma & Weenink,

2016). Prior to measurement, accuracy of the target productions was assessed and it was verified that speakers produced the target correctly in both the first and third events of discourse-given trials. During acoustic annotation, the annotator was blind to the experimental condition of each trial.

Speech onset latencies (RTs) were calculated from the onset of the action animation (marked by the experimental software in a second acoustic channel synced to the speech) to the onset of the vowel of the determiner. Vowel onset for the determiner (rather than fricative onset) was chosen due to the difficulty in reliably identifying the onset of frication for /ð/. To obtain determiner vowel durations, vowel onsets were marked at a rising zero crossing when clear formant structure had emerged for the vowel and vowel offsets were marked at a rising zero crossing on/after a sharp drop in amplitude upon closure for an upcoming consonant. To obtain target noun durations, word onsets and offsets were marked at zero crossings. Phoneme- and class-specific criteria for noun onset and offset boundary marking are listed in Appendix B. Verb productions were measured but not analyzed, as they were not balanced across trials and were not designed to match across conditions for each target.

The reliability of acoustic measurements was assessed by having an additional phonetically-trained annotator measure 10% of the target trials (n = 192). These trials were randomly sampled from the entire set of trials selected for the noun duration analysis, with an equal number of trials sampled from each group, for each target word, for each experimental condition, and (as much as possible) from each participant. All measurements from each annotator were highly correlated (RTs: Pearson's $r = 0.899$; nouns: Pearson's $r = 0.954$; determiner vowels: Pearson's $r = 0.906$). The mean absolute difference was 27.1 ms between RT measurements, 19.7 ms between noun duration measurements, and 5.5 ms between determiner

vowel duration measurements.

2.2.5 Analysis

2.2.5.1 Duration measures. Three duration measures were considered for analysis: RT, determiner vowel duration, and target noun duration. The maximum number of observations available for analysis was 4,608 (48 participants produced 48 target words in 2 conditions). Trials were excluded from duration analysis if they were audibly disfluent (repetitions, false starts, elongations, pauses longer than 250 ms) at any point during the trial (N = 505, 11% of data) or if the speaker did not produce the correct target word (N = 688, 14.9% of data).

Separate linear mixed effects regressions were built for each of these duration measures using the lme4 package, version1.1-7 (Bates, Maechler, Bolker, & Walker, 2015), in R 3.2.4 (R Core Team, 2016). Following previous work with this paradigm (Kahn & Arnold, 2015), baseline models with a series of control factors were first built. After fitting the initial baseline model, control factors were included in the final baseline model only if they contributed significantly to model fit (assessed by testing the significance of each factor via model comparison; see Table 2.2 for final baseline model structure for each measure and Appendix C for results of control factors). Candidate variables for control factors were RTs, determiner durations, noun durations, lexical frequency (all continuous factors were log-transformed and centered), and block (contrast-coded; block 1 vs. block2). These models included the maximal random effects structure supported by the data, determined by building the maximal possible random effects structure (all possible random slopes for the by-item and by-participant intercepts) and simplifying until convergence was achieved. Baseline models included random intercepts for participants and items (target nouns) and random slopes for all significant control factors.

After fitting the baseline model, the fixed effects of interest were included. The dependent measure for each regressions was the log-transformed duration measure of interest. Fixed effects for these analyses included contrast-coded effects for group (L1 vs. L2) and discourse condition (discourse-given vs. discourse-new), and their interaction. These factors were also included in the random effects structure. The maximum random effects structure allowed by the data was included in all models (Barr, Levy, Scheepers, & Tily, 2013). Significance of main effects and interactions was assessed via nested model comparison.

2.2.5.2 Disfluencies. The rate of disfluencies produced by participants was considered in a separate analysis using a series of logistic mixed effects regressions. Trials excluded from duration analyses for both disfluency and naming errors were included in this analysis. The dependent variable for these analyses was a binary measure (disfluent vs. fluent). Fixed effects for these models included contrast-coded effects for group (L1 vs. L2) and discourse condition (discourse-given vs. discourse-new), as well as their interaction. Models included the maximum random effects structure supported by the data (Barr et al., 2013), including decorrelated random slopes for group, discourse condition, and block by item and decorrelated random slopes for discourse condition and block by participant. Nested model comparison was used to perform significance tests.

2.3 Results

2.3.1 Response Times

Observations lying more than 3 standard deviations from each participant's condition mean were excluded from analysis (N = 25, 0.5% of data). The model for response times (RTs) included decorrelated random slopes for block, determiner duration, noun duration, group, discourse condition, and the interaction between group and discourse condition by item. For by

participant random effects, decorrelated random slopes for block, determiner duration, noun duration, and condition were included. Models were refit after removing observations with greater than 2.5 standardized residuals in the original model (N = 76, 1.6% of data).

Table 2.2. Summary of regression results (direction of effects) for control factors in the main analyses. Rows correspond to control variables considered for inclusion in models for each dependent measure (in columns). Direction of effects only shown for significant and marginal effects. Italicized text indicates a marginal effect. Grey indicates the variable was not included in the model.

	Response time (RT)	Determiner vowel	Target noun
Block	Positive	*Positive*	Negative
Log frequency			Negative
Log RT		Positive	Positive
Log determiner duration	Positive		
Log noun duration	Positive	Positive	

Table 2.3. Summary of regression results (direction of effects) for factors of interest in the main analyses. Rows correspond to variables considered for inclusion in models for each dependent measure (in columns). Direction of effects only shown for significant and marginal effects. Italicized text indicates a marginal effect.

	Response time (RT)	Determiner vowel	Target noun
Group	Negative	n.s.	Negative
Discourse condition	Negative	Negative	Negative
Group X discourse condition	n.s.	*Negative* L1: negative L2: n.s.	n.s.

Overall, L2 speakers were slower to initiate speech than L1 speakers (L1: mean = 1833.60 ms, SE = 82.30 ms; L2: mean = 2117.23 ms, SE = 81.83 ms; $\beta = -0.13$, SE = 0.06, $\chi^2(1) = 4.91$, $p < 0.05$). Similarly, response times were longer in discourse-new compared to discourse-given conditions (discourse-given: mean = 1918.08 ms, SE = 61.88 ms; discourse-new: mean = 2037.05 ms, SE = 69.07 ms; $\beta = -0.04$, SE = 0.01, $\chi^2(1) = 24.43$, $p < 0.001$), indicating words

with high discourse-dependent probability were easier to plan and initiate. However, the difference in response times across discourse conditions was similar across groups (β = 0.02, SE = 0.01, $\chi^2(1) = 2.18$, $p > 0.05$). These results are shown in Figure 2.2.

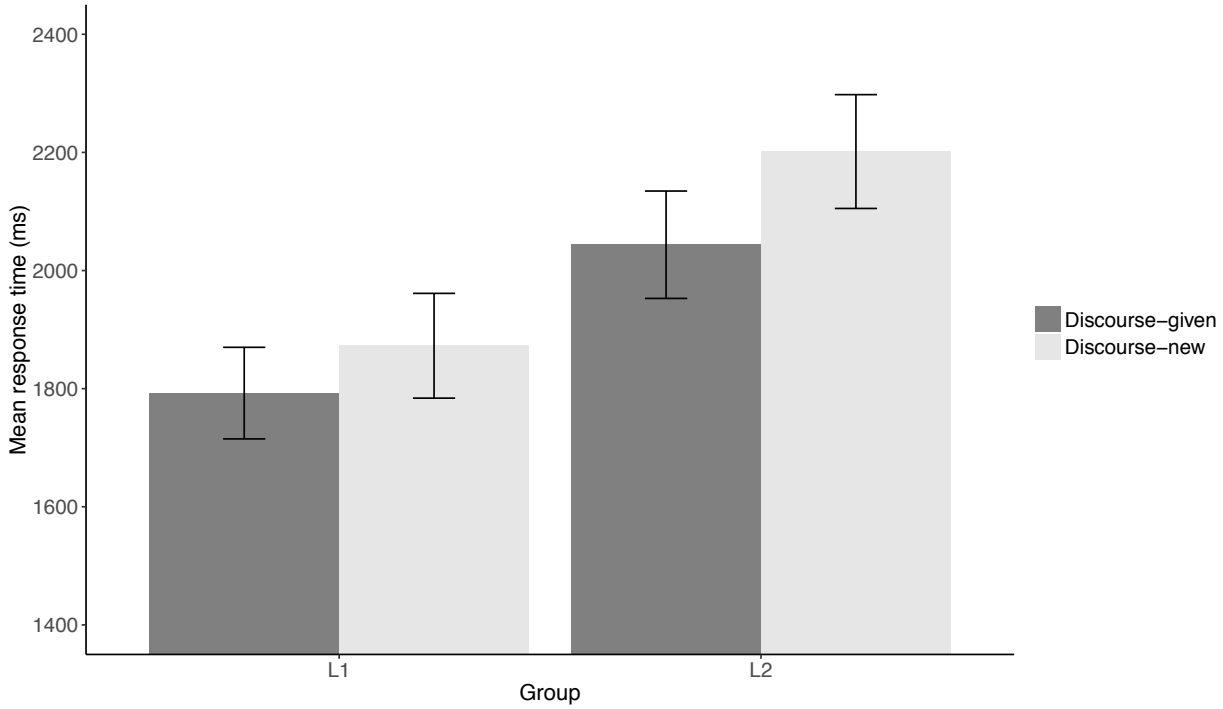

Figure 2.2. Mean response time (ms) across groups and discourse conditions. Error bars show standard error.

Together, the results of this analysis replicate a series of previous findings. As in Kahn and Arnold (2015), L1 speakers exhibited significantly longer planning times (indexed by RTs) in discourse-new compared to discourse-given conditions. Furthermore, these results indicate delays in speech planning for L2 vs. L1 speakers. This bilingualism-related slowing has been well-attested in the literature for bare picture naming tasks (e.g., Gollan, Montoya, Fennema-Notestine, & Morris, 2005) and as well as picture naming in sentence generation tasks similar to the current task (Runnqvist, Gollan, Costa, & Ferreira, 2013). However, despite this overall processing slow-down, and a numerical decrease in discourse-dependent effects, L2 speakers exhibited no significant deficits in how probabilistic information influenced processing; similar

facilitatory effects were found for L1 and L2 speakers in sentences containing a target noun with high discourse-dependent probability. This result stands in contrast to existing findings of differences between L1 and L2 speakers in the influence of other types of probabilistic information on processing (namely, influence of lexical frequency; Gollan et al., 2011).

2.3.2 Noun Durations

Observations lying more than 3 standard deviations from each participant's condition mean were excluded from analysis (N = 7, 0.2% of data). Furthermore, it is important for analysis of noun durations to be within participant and within item, as word durations vary greatly across individual speakers and words. Therefore, observations in the noun duration dataset were matched within participant such that an observation was only included if the target item was produced (following disfluency and outlier trimming) for both discourse conditions (N = 598 excluded; 13% of data). Observations were further matched across groups by randomly excluding L1 observations until each target item was represented equally across groups (N = 574 excluded; 12.5% of data). This matching across groups was important to ensure that any differences across groups could not be due to imbalances in the individual words included in the analysis (i.e., many observations of certain words in the L1 group vs. few for the L2 group). After these exclusions, a model for noun durations was built that included decorrelated random slopes for block, RT, group, discourse condition, and the interaction between group and discourse condition by item. For by-participant random effects, decorrelated random slopes for block, RT, frequency, and condition were included. Model-based outlier trimming was not performed for this analysis, as it would require re-matching the observations across items and groups.

Overall, L2 speakers produced longer word durations than L1 speakers (L1: mean =

335.27 ms, SE = 8.36 ms; L2: mean = 409.23 ms, SE = 18.59 ms; $\beta = -0.18$, SE = 0.05, $\chi^2(1) = 12.63$, $p < 0.001$). Furthermore, speakers produced significantly shorter word durations in discourse-given compared to discourse-new conditions (discourse-given: mean = 357.55 ms, SE = 11.64 ms; discourse-new: mean = 386.97 ms, SE = 11.46 ms; $\beta = -0.07$, SE = 0.01, $\chi^2(1) = 43.61$, $p < 0.001$). However, the magnitude of the effect of discourse condition on word durations did not differ across groups, indicated by a non-significant group by discourse condition interaction ($\beta = 0.008$, SE = 0.02, $\chi^2(1) = 0.28$, $p > 0.05$). [4]

[4] Three L2 speakers had substantially longer overall word durations than the rest of the group (230.6 ms difference), leading to substantially higher variability in L2 word durations (shown in Figure 2.3). To ensure that our results were not driven by these three speakers, the noun duration analyses were replicated excluding these speakers along with the last three L1 speakers who participated in the study. The same observation matching procedures were followed for this analysis, and a scaled-down Monte Carlo simulation was performed to control for differences in the observation sampling. Across 10 random samples, the group and discourse condition effects were significant 100% of the time, while the group by discourse condition interaction was never significant.

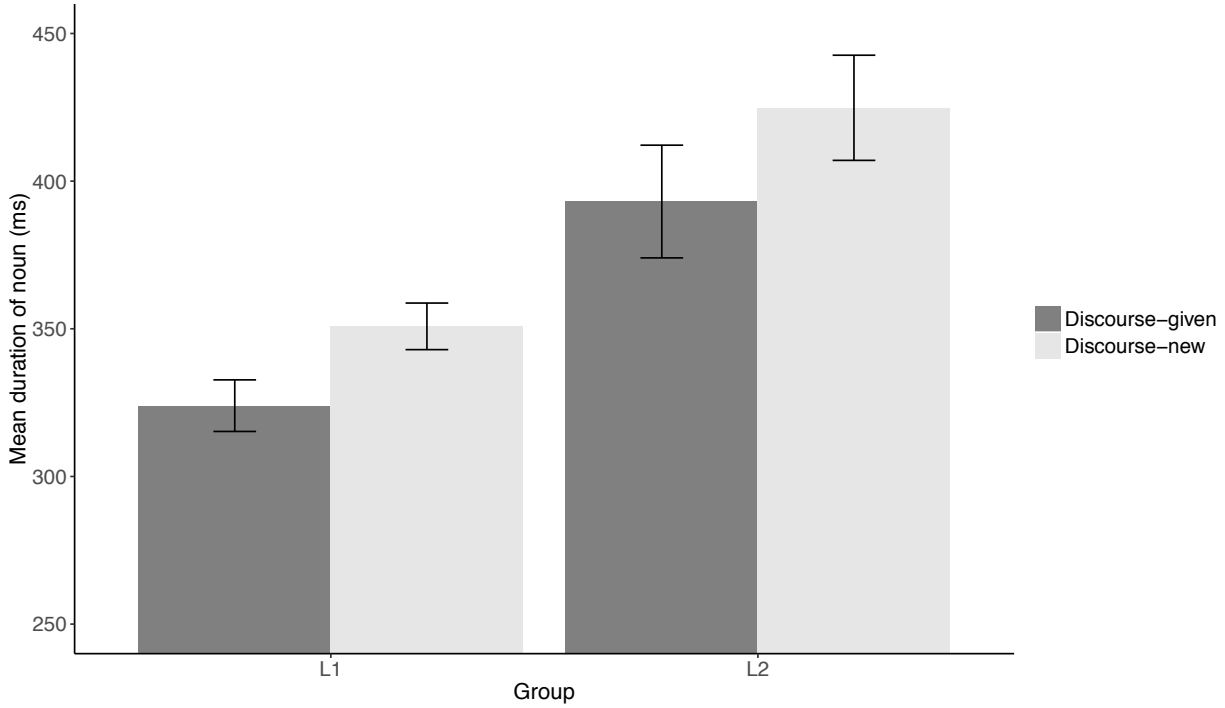

Figure 2.3. Mean duration of nouns (ms) across groups and discourse conditions. Error bars show standard error.

These results reveal that L1 and L2 speakers do not differ in the magnitude of discourse-dependent probabilistic reduction produced on target noun durations (see Figure 2.3). However, substantial differences in variability in word durations across groups raise the possibility that the lack of interaction could be driven by differences across groups in the set of words included in the analysis. To investigate this possibility, a Monte Carlo simulation was performed. The random sampling undertaken to match observations across groups (see above) was repeated 1000 times. The model built for analyzing word durations was re-run on each of these 1000 samples (the models failed to converge an additional 42 times). The result of interest for this simulation was the distribution of p-values across samples; for each effect in the model, we considered the proportion of p-values that passed the threshold of significance (i.e., $p < 0.05$).

The results of this simulation revealed that the majority of effects held across the random samples. In particular, among the control factors, the effects of log frequency and RT replicated

(frequency: 99.3% significant samples; RT: 100% significant samples). However, the effect of block on word durations was not reliable across samples (22.6% significant samples). Among the critical effects of interest, all results replicated; the effects of group and discourse condition on word durations were reliable (100% significant samples for each), while the group by discourse condition interaction never achieved significance across samples (0% significant samples). These results reveal that the lack of group by discourse condition interaction could not be attributed to differences in the words represented in the L1 vs. L2 dataset.

As with RTs, the noun duration results are consistent with existing findings in the literature. A number of other studies have also found that nouns are significantly reduced in duration when discourse-given compared to discourse-new (e.g., Kahn & Arnold, 2015). Existing studies of L2 speech production have also found that L2 speakers produce overall longer word durations than L1 speakers (Baker et al., 2011; Guion et al., 2000). The current results also extend previous findings that L2 speakers show comparable levels of reduction of nouns when discourse-given vs. discourse-new in reading tasks (Baker et al., 2011) by demonstrating this effects hold in a more challenging picture naming task, which requires semantic processing.

2.3.3 Determiner Vowel Durations

Observations lying more than 3 standard deviations from each participant's condition mean were excluded from analysis (N = 11, 0.2% of data). Furthermore, observations were excluded when the vowel of the determiner was devoiced, as the duration of the vowel was difficult to measure reliably (N = 191, 4.1% of data). In contrast to the noun analysis, there was no concern about imbalances in word representation across groups (as this analysis includes only one word). Therefore, random sampling to match observations across groups was not done for

determiners.[5] The model for determiner vowel durations included decorrelated random slopes for block, RT, noun duration, group, discourse condition, and the interaction between group and discourse condition by item. For by participant random effects, decorrelated random slopes for block, RT, noun duration, and condition were included. Models were refit after removing observations with greater than 2.5 standardized residuals in the original model (N = 64, 1.4% of data). Control factors and fixed effects of interest included in the model are summarized in Table 2.2.

The duration of determiner vowels did not differ significantly across groups (L1: mean = 43.26 ms, SE = 2.55 ms; L2: mean = 48.75 ms, SE = 2.89 ms; $\beta = -0.04$, SE = 0.08, $\chi^2(1) = 0.30$, $p > 0.05$). There was a main effect of discourse condition on determiner vowel durations (discourse-given: mean = 44.78 ms, SE = 1.99 ms; discourse-new: mean = 47.27 ms, SE = 1.95 ms; $\beta = -0.03$, SE = 0.01, $\chi^2(1) = 6.82$, $p < 0.01$), indicating that determiners modifying discourse-given nouns were significantly shorter than those modifying discourse-new nouns. However, this main effect was modulated by a marginal interaction with group ($\beta = -0.05$, SE = 0.02, $\chi^2(1) = 3.80$, $p = 0.051$). Follow-up regressions revealed a significant main effect of condition for L1 speakers (discourse-given: mean = 41.70, SE = 2.42; discourse-new: mean = 44.87, SE = 2.75; $\beta = -0.05$, SE = 0.02, $\chi^2(1) = 9.16$, $p < 0.01$) but not L2 speakers (discourse-given: mean = 47.87, SE = 3.07, discourse-new: mean = 49.68, SE = 2.74; $\beta = -0.003$, SE = 0.02,

[5] Given the high sensitivity of function word duration to probabilities of surrounding words (e.g., the target noun), one possibility worth considering is that certain nouns might be more or less effective at inducing reduction of the preceding determiner. If this relationship existed, it would motivate adoption of the same random sampling procedure for determiners as was done for the nouns. However, given that the frequency of the preceding noun had no influence on determiner duration (it did not contribute significantly to fit of the baseline model), it is unlikely that such an effect accounts for our results.

$\chi^2(1) = 0.03$, $p > 0.05$). These results indicate that L1, but not L2, speakers reduce the duration of determiners modifying discourse-given nouns compared to those modifying discourse-new nouns. L2 speakers' vowel durations in both conditions are roughly equivalent to L1 speakers' durations of determiners preceding discourse-new nouns (see Figure 2.4).

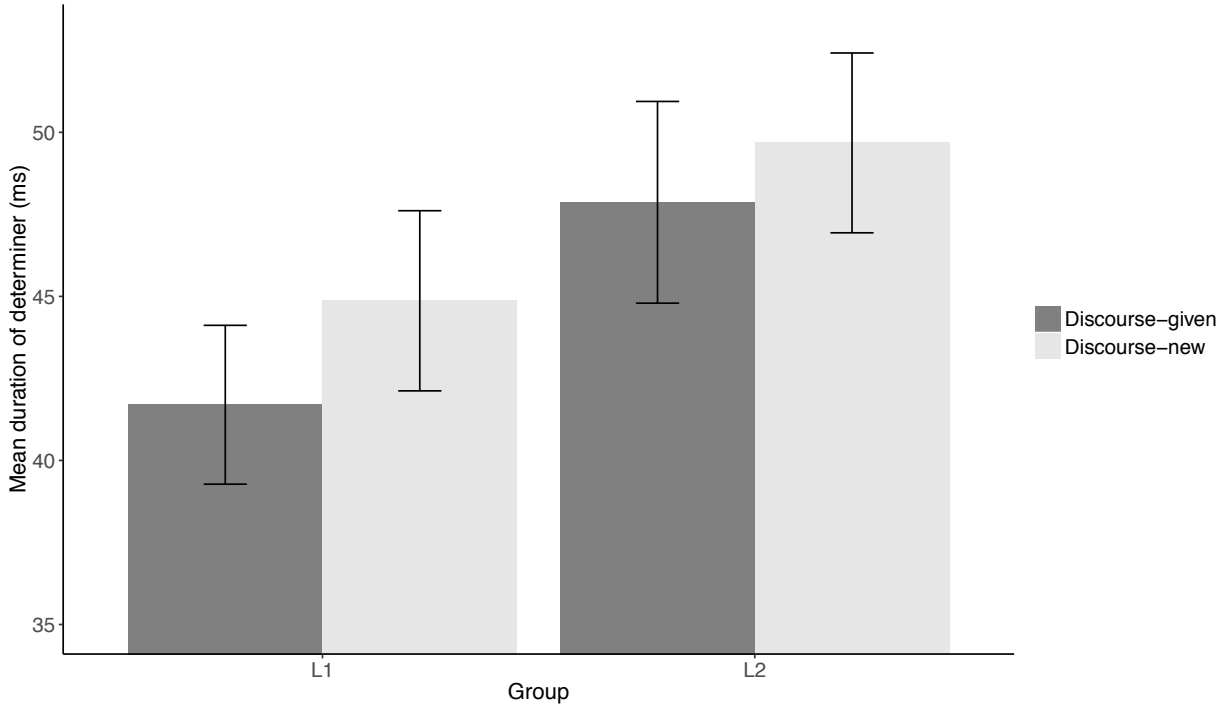

Figure 2.4. Mean duration of determiner vowels (ms) across groups and discourse conditions. Error bars show standard error.

These results are consistent with previous findings. As in Kahn and Arnold (2015), L1 speakers produce significantly shorter determiners in sentences with a discourse-given vs. discourse-new target noun. Furthermore, while Baker et al. (2011) observed that L2 speakers produced significantly longer function words than L1 speakers, this result did not seem to hold for the most common function words, such as the definite determiner *the*. The current results confirm this interpretation of their results; L2 speakers did not produce significantly longer determiners overall compared to L1 speakers (with durations roughly equivalent to the L1 determiners preceding discourse-new nouns). Interestingly, despite showing no general deficit in

the production of determiners, L2 speakers exhibit a particular deficit in how probabilistic information influences determiner production. In particular, L2 speakers did not reduce determiners in discourse-given vs. discourse-new conditions as L2 speakers did, despite having done so to an L1-like degree for RTs and noun durations. This set of results suggests that the planning and execution of speech differs substantially as a function of linguistic experience, particularly in how probabilistic information influences processing. Further discussion of the interpretation of these results, including sources of the deficit in how probabilistic information influences processing, will be explored in the general discussion.

2.3.4 Disfluencies

The results of the logistic mixed effects regression revealed that there was a main effect of group on disfluencies, where L2 speakers produced a significantly higher proportion of disfluencies compared to L1 speakers (L1: mean = 37.07, SE = 2.02; L2: mean = 14.71, SE = 1.49; $\beta = 1.53$, SE = 0.22, $\chi^2(1) = 38.64$, $p < 0.001$). Speakers also produced a significantly higher proportion of disfluencies in discourse-new vs. discourse-given conditions, indicated by a main effect of discourse-condition (discourse-given: mean = 23.52, SE = 1.93; discourse-new: mean = 28.26, SE = 2.30; $\beta = 0.34$, SE = 0.08, $\chi^2(1) = 15.25$, $p < 0.001$). As the experiment progressed, speakers produced lower proportions of disfluencies, shown by a main effect of block (block 1: mean = 28.17, SE = 2.34; block 2: mean = 23.61, SE = 1.88; $\beta = -0.29$, SE = 0.08, $\chi^2(1) = 11.59$, $p < 0.001$). None of the two-way interactions were significant (all $\chi^2(1) < 2$, ps > 0.05). There was a trend towards a three-way interaction ($\beta = -0.52$, SE = 0.32, $\chi^2(1) = 2.56$, $p < 0.11$), driven by a block by discourse condition interaction for L2 vs. L1 speakers (for the former group, the effect of discourse condition was stronger in the first block).

Analysis of disfluencies largely parallels the results in fluent speech. In particular, L2

speakers were both more disfluent, were slower to initiate speech, and produced longer fluent word durations on average. Similarly, speech was more likely to be disfluent, be more difficult to initiate, and have longer durations on average in discourse-new compared to discourse-given conditions. Finally, parallel to the majority of duration analyses, L2 speakers did not produce significantly more disfluencies in discourse-new conditions compared to L1 speakers. These results suggest that the processing of probabilistic information proceeds largely similarly across groups (at least for content words), both during successful speech processing and when it breaks down, regardless of differences in linguistic experience.

2.3.5 Post-Hoc Analysis: Lexical Frequency

An additional analysis considered the influence of lexical frequency on the magnitude of the condition effect on word durations (Baker & Bradlow, 2008) and explored whether this type of probabilistic information had differential influence on L1 vs. L2 speakers (e.g., Baker et al., 2011; Sadat, et al., 2012). Due to the strong negative correlation between word length and word frequency (i.e., Zipf's law; Zipf, 1949), it is critical to control for the effect of word length when considering the effect of lexical frequency on noun durations; unsurprisingly, longer words have longer durations. Therefore, models for this analysis included an additional control factor for word length in phonemes (centered). A by-subject random slope for length was added to the model. Lexical frequency (frequency per million from the SUBTLEX-US corpus; Brysbaert & New, 2009) was included as a continuous (log transformed and centered) fixed effect and in interactions with group and discourse condition. Models for this analysis were otherwise identical to those from the main noun duration analysis.

As in the main analysis, there was a significant main effect of group ($\beta = -0.17$, SE = 0.05, $\chi^2(1) = 10.86$, $p < 0.001$) and a significant main effect of discourse condition ($\beta = -0.08$, SE

= 0.01, $\chi^2(1) = 43.16$, $p < 0.001$). While there was a significant influence of word length on noun durations ($\beta = 0.09$, SE = 0.02, $\chi^2(1) = 15.15$, $p < 0.001$), the main effect of frequency on word durations was marginal ($\beta = -0.04$, SE = 0.02, $\chi^2(1) = 3.69$, $p = 0.055$), providing weak evidence in favor of a frequency effect independent of length. Importantly, none of the interactions involving frequency reached significance (all $\chi^2(1) < 2.01$, ps > 0.05), confirming that the absence of the frequency main effect was not due to differences in the size of the effect across groups or conditions.

Due to the fact that this analysis implemented random sampling of items across groups, as in the main noun duration analysis, a MC simulation was performed to ensure that absence or presence any effects could not be due to the items represented in the random sample. One thousand random samples were drawn in the simulation and a model was fit to each sample dataset (an additional 42 models did not converge). The relevant statistic for this analysis is the proportion of samples that achieved significance below the $p < 0.05$ threshold.

As in the main analysis, the main effect of group, condition, and RT replicated across samples (100% significant samples for each). Also similar to the main analysis, the main effect of block replicated unreliably across samples (22% significant samples). The main effect of word length replicated reliably (100% significant samples), while the main effect of frequency on noun durations rarely achieved significance (3.2% significant samples). However, the frequency effect was at least marginally significant ($p < 0.1$) in 64.6% of the samples. Finally, none of the two-way interactions ever reached significance (0% significant samples), and the three-way interaction achieved significance in only 2.1% of samples.

Contrary to findings by Baker and Bradlow (2009), we observed no evidence that discourse-dependent probability interacts with lexical frequency to influence word durations.

Specifically, we did not observe enhanced reduction of high frequency compared to low frequency words in discourse-given conditions. It is perhaps unsurprising that this effect fails to replicate given that Baker and Bradlow's effect was driven largely by reduction of the highest frequency word in their corpus, which happened to be the function word *and*. The authors note that the correlation between discourse-dependent probability and lexical frequency disappears with this item removed. Considering these results together, we have growing evidence that these distinct probabilistic forces (i.e., discourse-level and language-level) do not interact during speech production to influence word durations. Furthermore, consistent with previous work by Baker et al. (2011) and Sadat et al. (2012), we observed that L1 and L2 speakers produced frequency effects of similar magnitudes. Combined with other results of the current study, this provides evidence for the hypothesis that probabilistic information influences content word processing (of any kind) similarly during L1 and L2 speech production.

2.4 General Discussion

The current study investigated differences in how probabilistic information – specifically, discourse-dependent probability – affects planning and articulation during L1 and L2 speech production in a demanding referential communication task. For RTs and production of content words, we asked whether L2 speakers show deficits in how probabilistic information influences processing; that is, do they produce different levels of probabilistic reduction on RTs and content words when compared to L1 productions? The results showed no such deficits. Both L1 and L2 speakers produced significantly shorter RTs and word durations in discourse-given vs. discourse-new conditions, and, critically, there was no significant difference in the size of the reduction effect across groups. A contrasting pattern was observed for function words. While L1 speakers reduced vowels in function words preceding discourse-given nouns, L2 speakers showed no

significant reduction effect. This suggests that the difficulties associated with L2 speech production prevent determiners from inheriting the discourse-dependent probability of the following nouns.

2.4.1 L2 Processing of Probabilistic Information during Noun Production

The effects for noun reduction are consistent with existing research, L2 speakers were significantly slower to initiate speech (e.g., Gollan et al., 2008) and produced significantly longer word durations in comparison to L1 speakers (Baker et al., 2011; Guion et al., 2000). These results were also consistent with previous findings with read speech, where L2 speakers produced discourse-dependent probabilistic reduction of the same degree as L1 speakers (Baker et al., 2011).

While the lack of differences in effect sizes for word durations is consistent with previous work, the lack of differences in effect sizes for RTs stands in contrast to a body of existing work that has demonstrated deficits for probabilities conditioned on a word's distribution within a language (e.g., lexical frequency; Gollan et al., 2008; Gollan et al., 2011; although, see Sadat et al., 2012 for conflicting results). These diverging results highlight the influence of linguistic experience on how probabilistic information impacts speech planning. For lexical frequency to influence L2 planning in a similar manner as L1 planning, L2 speakers must have similar experience as L1 speakers with the words in the language. As the frequency lag hypothesis (Gollan et al., 2008) argues, L2 speakers cannot have the same experience as L1 speakers. However, for discourse-dependent probabilities to influence planning in an L1-like manner, L2 speakers must only have the same experience of a word within a discourse as an L1 speaker. Our results indicate that L2 speech processing is not so demanding as to prohibit L2 speakers from tracking the discourse-dependent probability of a word, so speakers in each group have the same

experience. Therefore, we observe no differences between L1 and L2 speakers in the magnitude of discourse-dependent probabilistic reduction on RTs.

Examination of frequency effects in the post-hoc analysis provided additional insights into L2 processing of probabilistic information during noun production. In contrast to RT studies, previous studies investigating the influence of lexical frequency on noun durations have reported no significant or merely marginally significant differences in frequency effects on durations between L1 and L2 speakers (Baker et al., 2011; Sadat et al., 2012). The present results for noun durations are in line with these existing findings. Future research should consider whether this dissociation in measures of planning (RTs) vs. execution (durations) can be attributed to differences in the bilingual populations tested across studies. Studies by Gollan and colleagues (2008, 2011) include switched dominance bilinguals, who participated in their dominant L2, Sadat et al. tested highly balanced bilinguals in their L1, while Baker et al. and the current study elicited speech from unbalanced bilinguals in their L2. A limitation of these studies is that the frequency counts used to design the stimuli were derived from monolingual L1 corpora, which may not reflect differences across groups in speakers' experience with words in the language.

If differences across bilingual populations cannot account for differences across studies, it instead could be the case that frequency influences different stages of processing in distinct ways. Frequency lag effects on RTs are typically concentrated on the lower frequency items; L2 speakers are slower to plan speech overall, but especially for low frequency words (Gollan et al., 2011). One possibility is that overall L2 slowing in articulation puts speakers at a floor level of slowing; observing a frequency lag effect on durations would require even longer word durations, perhaps past an acceptable threshold (especially for bilingual speakers in Sadat et al.'s

study).

2.4.2 L2 Processing of Probabilistic Information during Determiner Production

For function word production, we observed a different pattern of results. L2 speakers did not produce significantly longer determiner vowel durations than L1 speakers, with L2 means in the range of determiners preceding discourse-new nouns in the L1. This is consistent with the observation that L1 and L2 speakers produce similar durations for the highest frequency function words in Baker et al. (2011). This result indicates that L2 speakers have sufficient experience with determiners, and their implementation as prosodic clitics, to produce determiner vowel durations that are similar to those in the L1. Interestingly, this particular set of L2 speakers have no experience with determiner production in their L1 (Mandarin), which one might predict would lead to generalized deficits in determiner production in the L2. Our results suggest that intensive experience with this high frequency function word allows the L2 speakers to overcome the lack of L1 base knowledge.

Despite having produced determiner vowels within the L1 range, as well as their success in deploying probabilistic information for content words, the L2 speakers failed to match the L1 speakers in reducing determiner vowels in discourse-given vs. discourse-new conditions. This function word-specific deficit is expected under the hypothesis that L2 speakers have a reduced scope of planning compared to L1 speakers (i.e., single words planned in sequence vs. multiple words in parallel), perhaps due to the high demand associated with L2 speech processing. This reduced scope would make it difficult for L2 speakers to use properties of the upcoming noun, such as its discourse-dependent probability, to modulate determiner vowel duration. Future research should investigate the reduced scope hypothesis further.

An alternative possibility is that L2 speakers have a reduced scope of planning due to the

lack of determiners in Mandarin, rather than due to increased demand. Future research could test these contrasting hypotheses by eliciting speech from L2 speakers who have L1 experience with determiners. Some support for this hypothesis comes from Schertz and Ernestus (2014), who found that Czech speakers (who use a determiner-like demonstrative pronoun in their L1) but not Norwegian speakers (who mark definiteness using suffixation in their L1) reduced English determiners based on the probability of the following noun. Furthermore, for L1 speakers of so-called late selection languages (e.g., Romance languages) determiner selection occurs late in noun phrase processing (Miozzo & Caramazza, 1999). While the scope of planning for such speakers may be similar as for speakers of English (not a late selection language), divergences in the time-course of determiner processing may lead to differences in the influence of probabilistic information.

One limitation of the current study, shared with Schertz and Ernestus (2014), is that determiners inherit probability only from nouns. Follow-up studies should investigate whether probability inheritance occurs with other types of function and content words, as well as within other types structural relationships (i.e. function words in adjunct vs. modifier position). Given the highly dependent structural relationship between determiners and nouns, and the fact that the planning of determiners depends on noun selection (Bock & Levelt, 1994), this probability inheritance could be restricted to this particular case. However, studies have also shown that other function words are sensitive to the probability of surrounding words (e.g., Bell et al., 2003). Future work could systematically investigate these potential differences.

The determiner duration results also provide important constraints on theories of how discourse-dependent probabilities influence processing to bring about reduction. A pure facilitation-based account of this reduction (e.g., Kahn & Arnold, 2012) predicts that both nouns

and determiners should show equivalent effects across L1 and L2. Both groups of speakers utilize the same production-based mechanisms that benefit from repeated planning, so both L1 and L2 speakers should therefore show discourse-dependent reduction. This cannot account for the current set of results, in which RTs and nouns benefit from facilitation but determiners do not. Facilitation of production mechanisms due to repetition of the target noun should impact both production of determiners and nouns (as the mechanisms are shared; Bürki et al., 2014). In contrast, the probability inheritance hypothesis can accommodate these findings. Under this account, discourse-dependent reduction of function words occurs because they inherit the probability of the nouns they modify. Due to the high demand of L2 processing overall, this probabilistic information is not inherited, so function word processing cannot be affected by discourse-dependent probabilistic information.

A facilitation-based account of reduction (Kahn & Arnold, 2012) could still account for these results by assuming that facilitation occurs over pre-defined chunks of words rather than for production mechanisms as a whole. Under this modified account, production of a determiner-noun chunk would be facilitated in discourse-given vs. discourse-new trials. We may, then, be able to account for the results if we then hypothesize that the chunk does not exist for L2 speakers (i.e., facilitation occurs separately for determiners and nouns) but it does for L1 speakers (i.e., facilitation occurs in tandem). This would predict that the recent production of a noun facilitates the chunk for L1 speakers, leading to reduction of both the determiner and noun, but only facilitates the noun for L2 speakers, leading to reduction of only the noun. Even with this modification, it is unclear why RT reduction should occur when determiner reduction does not. Therefore, we argue that the probability inheritance hypothesis is a more parsimonious account of these results, which fit in nicely with existing findings that function word reduction is

sensitive to the probability of surrounding words (Bell et al., 2003; Bell et al., 2009).

2.4.3 Probabilistic Information for the Speaker vs. for the Listener

Researchers have proposed a variety of mechanisms to account for how probabilistic information influences planning and articulation, including both speaker-based and listener-based processes. According to speaker-based theories, the probabilities associated with a word directly influence the mechanisms underlying speech production, either in terms of how probabilities are stored, how they influence lexical access, or both. Storage-based accounts argue that the probabilities associated with a word are encoded in (long-term) linguistic representations (via resting activation (Dell, 1990); via phonetically-specified representations (Pierrehumbert, 2001, 2002; Seyfarth 2014)). These representations, and the probabilities they encode, influence processing, and lead to reduction during planning and articulation. For example, retrieval of high probability lexical items may be easier than for low probability items due to high resting activation in long-term representations. Words that are easy to retrieve will be selected more quickly and will be easy to articulate, leading to hypoarticulation. Ease of retrieval has also been attributed to temporary boosts in resting activation due to priming from recent retrieval (Arnold & Watson, 2015; Fraundorf, Watson, & Benjamin, 2014; Kahn & Arnold, 2012), although the current results rule out this possibility. While these storage-based and processing-based accounts are difficult to dissociate, they both reflect purely speaker-based mechanisms.

In contrast, listener-based theories argue that the probabilities associated with a word reflect some calculation undertaken by the speaker for the benefit of the listener. The main assumption underlying such theories is that there is some communicative intent of the articulatory reduction resulting from the influence of probabilistic information on processing. Lindblom's (1990) Hyperspeech-Hypospeech theory proposed that speakers hyperarticulate

words that they think will be difficult for their listener to understand (e.g., low probability words), and reduce words that they think will be easy (e.g., high probability words), thus balancing the need for effective communication and the desire to minimize articulatory effort. Information theoretic accounts, such as the Smooth Signal Redundancy Hypothesis (Aylett & Turk, 2004) and the Uniform Information Density Hypothesis (Levy & Jaeger, 2007), take a similar approach. These theories contend that effective communication requires a trade-off between a desire to produce brief (i.e., reduced) linguistic signals and the need for communication error to be low (Aylett & Turk, 2004; Levy & Jaeger, 2007; Pate & Goldwater, 2015). However, highly reduced signals could impede successful communication if "noise" enters the system (Shannon, 1948) on the speaker side (e.g., speech errors), in the environment (e.g., a noisy room), or on the listener side (e.g., distraction). Therefore, brief signals are optimal only when the message has high probability, because such messages will have intrinsically low probability of communication error. By contrast, communication of messages with low probability should be encoded with relatively longer signals to avoid errors in communication.

Either class of theory can account for the results of the current study. Following speaker-based theories, the discourse-dependent probability of a word could be stored in short-term representations. Encoding of this high probability causes a temporary boost in the resting activation for the representation, which facilitates lexical retrieval and leads to reduction. Under listener-based theories, speakers know they can reduce word with high discourse-given probability without risking perception error on the part of their (potential) listener. However, it is likely that both speaker-driven and listener-driven forces underlie these effects (e.g., Arnold, Kahn, & Pancani, 2012; Rosa, Finch, Bergeson, & Arnold, 2015). Our results with L2 speakers indicate that whichever of these mechanisms drives the influence of discourse-dependent

probability on reduction for L1 speakers also does so for L2 speakers, at least during the planning and execution of content words. During function word production, we argue that determiners do not have high discourse-dependent probability at all for L2 speakers (for the reasons discussed above), and thus they would not be expected to reduce the determiners under any of these theories.

An underlying assumption to the listener-based theories discussed above is that listeners are (implicitly) aware of the relationship between probability and reduction (Mitterer & Russell, 2013); listeners should know that reduced words are likely to have high probability, and reduced words have low probability. That is, when listeners hear a reduced syllable, they predict that the syllable belongs to word with high vs. low probability according to some factor, such as discourse status (Arnold, 2008; Dahan, Tanenhaus, & Chambers, 2002). Listeners also use acoustic information earlier in the sentence (e.g., at the determiner) to make predictions about upcoming words (e.g., Arnold, Tanenhaus, Altmann, & Fagnano, 2004). While L1 listeners undoubtedly use discourse-dependent reduction to make predictions during speech production, it is unclear if L2 listeners can. Under certain circumstances, L2 listeners cannot make predictions at all (e.g., Lew-Williams & Fernald, 2010). However, if L2 processing of probabilistic information leads to L1-like reduction, as in the current study, it is possible L2 listeners will also be able to use this reduction as a predictive cue during speech perception (Pickering & Garrod, 2013). Study 2 investigates this question.

2.5 Conclusions

In conclusion, the current study sheds light on the role of linguistic experience in how probabilistic information influences speech production processing by L1 and L2 speakers of English. These results provide important evidence for understanding how probabilistic

information influences L2 speech planning and execution. The L1-L2 comparison also reveals important insights into how the linguistic system works more generally. By leveraging differences between these groups of speakers, and differences between content and function word production, we were able to distinguish between competing theories for the mechanism underlying the influence of probabilistic information on processing. This work highlights the utility of bilingual research for understanding both monolingual and bilingual language processing.

CHAPTER 3

3.1 Introduction to Study 2

No single utterance of a word has identical acoustic-phonetic properties, even when produced by the same speaker. Such variations are widespread and stem from a number of sources, including the probability of a word occurring in a discourse. Discourse-given words have high discourse-dependent probability due to having been previously produced in the current discourse, while discourse-new words entering the discourse for the first time have low probability. Words with high probability are typically phonetically reduced (e.g., have shorter overall duration, have shorter and more centralized vowels, etc.) compared to those with low probability (e.g., Jurafsky, Bell, Gregory, & Raymond, 2001), and this principle applies to discourse-dependent probability as well (with discourse-given words reduced relative to discourse-new; e.g., Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Fowler & Housum, 1987). Recent investigations have also shown that speakers reduce not only nouns with high discourse-dependent probability but also the preceding determiners (Kahn & Arnold, 2012, 2015; Study 1).

Listeners are aware of this relationship between discourse-dependent probability and reduction (on content words) and exploit it during speech perception to predict the identity of words before receiving complete input (Arnold, 2008; Dahan, Tanenhaus, & Chambers, 2002; Isaacs & Watson, 2010). For example, when there are two phonologically-related candidates for the current (incomplete) input (e.g., *candy* and *candle* after hearing /kæn…/), listeners associate reduced input with a discourse-given candidate and unreduced input with a discourse-new candidate. This discourse-dependent reduction conveys useful information for the listener about

the identity of words, potentially allowing the listener to more quickly recognize words as the speech stream continuously unfolds.

Studies have also shown that listeners use information earlier in the sentence (e.g., at the determiner preceding a target noun) to make predictions about the identity of upcoming words. For example, when there is a disfluency prior to a target, listeners predict that the target will be discourse-new (as disfluencies occur more often in discourse-new contexts; Arnold & Tanenhaus, 2007) rather than discourse-given (Arnold, Tanenhaus, Altmann, & Fagnano, 2004). Prediction based on the determiner facilitates recognition of a word before listeners have received any bottom-up input about that word. Given that determiners undergo discourse-dependent reduction just as nouns do, this leaves the possibility that this reduction can serve as a predictive cue to the identity of the upcoming noun.

For listeners to use probabilistic information as a predictive cue during speech perception, they must have sufficient experience with the probability distributions underlying the phonetic variation in question. Listeners using their native language (L1 listeners) have this experience, and should have little difficulty leveraging these distributions. However, listeners using their second language (L2 listeners) do not have the same experience with a language as L1 listeners, meaning the probability distributions they have experienced will likely not match the distributions in incoming L1 speech. Evidence suggests that such divergences in experience across L1 and L2 listeners may contribute to deficits in L2 speakers' ability to use knowledge specific to the L2 to facilitate language comprehension (e.g., Lew-Williams & Fernald, 2010). Such work suggests that L2 listeners may have difficulty utilizing discourse-dependent reduction relative to L1 listeners. However, L2 speakers reduce words with high discourse-dependent probability to a similar degree as L1 speakers (Baker, Baese-Berk, Bonnasse-Gahot, Kim, Van

Engen, & Bradlow, 2011; Study 1), which would suggest that L1 and L2 listeners have similar experience with these particular probability distributions. Therefore, an alternative possibility is that the increased demand associated with L2 processing may impede L2 listeners' ability to exploit any experience they do have (Kaan, 2014).

The current study investigates whether L1 and L2 listeners show similar use of probabilistic information as a predictive cue during speech perception. We examine both the ability to predict the identity of a noun before receiving complete input as well as the ability to predict an upcoming noun based on the presence vs. absence of reduction on a preceding determiner. We begin by briefly summarizing how probabilistic information influences processing during speech production, and then turn to existing research that has investigated influences of probabilistic information on online speech perception. We then review previous work on such influences in L2 speech perception, motivating the design of our current study.

3.1.1 Processing Probabilistic Information in Production

A single word can be associated with a series of probabilities conditioned on various factors or levels of linguistic structure. For example, lexical frequency can be thought of as the probability of a word within a particular language (independent of particular context within a language). Of interest for the current study is discourse-dependent probability, or the probability of a word conditioned on its distribution within some discourse. The processing of these types of probabilistic information has consistent influence on the phonetic properties words. Many studies have observed that high probability words are associated with phonetic reduction in the duration of the stressed vowel or entire word, spectral qualities of the word, as well as intensity (e.g., Aylett & Turk, 2004; Baker et al., 2011; Baker & Bradlow, 2009; Bell et al., 2009; Fowler & Housum, 1987; Gahl, 2008; Gahl & Garnsey, 2006; Kahn & Arnold, 2012, 2015; Lam &

Watson, 2010, 2014; Liberman, 1963; Munson & Solomon, 2004; Pate & Goldwater, 2015; Scarborough, 2010).

In studies that establish a discourse in the context of the experiment, such as event description tasks (e.g., Lam & Watson, 2010), results have shown robust reduction effects on the durations of words with high vs. low discourse-dependent probability (Baker et al., 2011; Baker & Bradlow, 2009; Bell et al., 2009; Fowler & Housum, 1987; Kahn & Arnold, 2012, 2015; Lam & Watson, 2010, 2014). Furthermore, similar effects of discourse-dependent probability on production of nouns have been observed for L1 and L2 speakers (Baker et al., 2011; Study 1), suggesting such effects are robust to differences in linguistic experience. However, this may not be the case for all word classes; while studies have shown that L1 speakers also reduce determiners in high probability conditions (Kahn & Arnold, 2012, 2015), a recent study failed to find the same reduction for L2 speakers (Study 1).

A number of mechanisms have been proposed to account for influences of probabilistic information on processing, including speaker-based and listener-based accounts. Under speaker-based theories, the probabilities associated with a word directly influence the mechanisms underlying speech production, either in terms of how probabilities are stored, how they influence lexical access, or both (Arnold & Watson, 2015; Fraundorf, Watson, & Benjamin, 2014; Kahn & Arnold, 2012). Such theories place no specific requirement on the capabilities of listeners. In contrast, listener-based theories argue that the probabilities associated with a word form the basis of an implicit calculation by the speaker of how production should be modified to benefit the listener. Multiple listener-based theories exist, with some differences in nuance, but with the common argument that high probability words can (or should) be reduced because they are easy for the listener, while low probability words should be unreduced to avoid listener error (Aylett

& Turk, 2004; Levy & Jaeger, 2007; Lindblom, 1990; Pate & Goldwater, 2015; see Pierrehumbert, 2001, 2002; Seyfarth, 2014, for theories in which listener-based preferences are stored in production mechanisms).

Critically, these listener-based theories assume that the covariation of probability and reduction is meaningful to the listener; that is, listeners are (implicitly) aware that reduced words are likely to have high probability, while unreduced words have low probability (Mitterer & Russell, 2013). This awareness should allow listeners to use reduction as a signal to the probability of a word, which then serves as a cue to the word's identity. As discussed in the next section, a number of studies have established that listeners use the relationship between discourse-dependent probability and reduction during speech perception (Arnold, 2008; Dahan et al., 2002; Isaacs & Watson, 2010).

3.1.2 Prediction in L1 Perception

Listeners are able to make use of probabilistic information during speech perception due to the continuous, incremental processing of incoming speech (McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008). As a listener encounters the unfolding speech signal, a number of candidate representations become active and compete for selection. For example, after hearing the syllable /kæn-/, listeners activate not just the word *can* but also other possible continuations such as *candy* or *candle*. Evidence in support of this type of incremental processing comes from visual world eye-tracking experiments where listeners' eye movements are recorded as they receive auditory input. At the point at which they have heard /kæn-/, listeners looked more to pictures whose names contained the syllable (i.e., the cohort competitors, *candy* and *candle*) than to phonologically unrelated pictures (e.g., *lemon* and *skunk*; Allopenna, Magnuson, and Tanenhaus, 1998).

When multiple candidates are consistent with the incomplete input, listeners may use probabilistic information, such as lexical frequency, in tandem with phonetic and phonological structure to predict the identity of the word. We adopt a broad definition of prediction, which allows both concurrent (i.e., co-occurring with bottom-up input associated with the target) and contextual (i.e., preceding bottom-up input associated with the target) information to influence processing of a target word (n.b., a stricter definition might limit prediction to only the latter case; e.g., Kuperberg & Jaeger, 2016). Under this broad definition, predictive processing can occur *during* target processing when listeners receive ambiguous input. For example, as discussed above, after hearing part of a target word, listeners activate cohort competitors as opposed to phonologically related properties. Listeners can also predict a word's identity by using probabilistic information about the word itself, such as lexical frequency, as a predictive cue. Dahan, Magnuson, and Tanenhaus (2001) found that listeners looked more and were faster to look to high vs. low frequency competitors (e.g., *candy* vs. *candle*); listeners predict that the high frequency word is the more likely candidate, all else being equal. Phonetic and prosodic information associated with the target can also drive predictive processing during speech perception (e.g., Arnold, 2008; Dahan et al., 2002; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Isaacs & Watson, 2010; Ito & Speer, 2008; McMurray, Tanenhaus, & Aslin, 2002; Mitterer & McQueen, 2009; Mitterer & Russell, 2003). Building on Dahan et al.'s (2001) findings, Mitterer and Russell (2003) demonstrated that listeners look more to high vs. low frequency competitors when presented with phonetically reduced productions. These findings indicate that listeners are sensitive not only to target-concurrent probabilistic information but also the relationship between probability and phonetic variation (e.g., reduction of high vs. low frequency forms).

Other types of target-concurrent probabilistic information, such as discourse-dependent probability, are conditioned on contextual factors. A series of studies have shown that listeners use the reduction associated with discourse-dependent probability to predict the complete form of words for which they have received partial, ambiguous input. These studies have framed reduction in terms of the processing of accented and unaccented words. Accentedness is typically signaled by the presence of a pitch accent, which consists of a large change in pitch and long duration on the stressed syllable (e.g., Ladd, 1996). Therefore, accented words are unreduced (large pitch change, long duration) while unaccented words are reduced (small pitch change, short duration). Research has shown that listeners have expectations about the relationship between word-level reduction and discourse-dependent probability. Bock and Mazzella (1983) found facilitated comprehension (faster response times, or RTs, to indicate comprehension of the sentence) when discourse-given words were reduced vs. unreduced and when discourse-new words were unreduced vs. reduced.

Other studies have shown that listeners exploit these expectations during online speech perception, using discourse-dependent probabilistic reduction as a predictive cue. In an eye-tracking study, Dahan et al. (2002) manipulated the discourse-dependent probability of target nouns and whether the noun was reduced (unaccented) or unreduced (accented). In discourse-given trials, listeners were presented with a series of instructions such as *Put the candle below the triangle… Now put the candle above the square*, where the target *candle* was discourse-given in the second (critical) instruction. In these discourse-given trials, listeners looked to the target more when reduced (expected) vs. unreduced (unexpected), and the competitor more when the target was unreduced vs. reduced. In trials where the target was discourse-new (heard for the first time in the second instruction), listeners showed the complementary pattern for competitor looks;

they looked to the competitor more when the target was reduced vs. unreduced. For target looks, they showed no difference (looking equally when reduced vs unreduced). These results demonstrated that listeners make predictions about (ambiguous) unfolding input based on the level of reduction associated with discourse-given vs. discourse-new words (see also Isaacs & Watson, 2010).

Arnold (2008) replicated Dahan et al.'s results with reduced words with both child and adult listeners. She demonstrated that listeners looked more to discourse-given vs. discourse-new images when the word was reduced, but found that neither group of listeners showed preferential looks to the discourse-given vs. discourse-new pictures when the production was unreduced. These results suggest that not all discourse-dependent probabilistic information generates predictions in the same way; reduced productions were interpreted by listeners as providing information about the discourse-dependent probability of the word, while unreduced productions were not. Arnold argues this is due to the fact that unreduced productions can felicitously signal both discourse-given and discourse-new status, while reduced productions are most consistent with a discourse-given target.

With our broad definition of prediction, we argue that probabilistic reduction on target nouns themselves can be used to predict the identity of an ambiguous target. However, prediction certainly comes into play, even under traditional definitions, when probabilistic information earlier in the utterance influences the processing of upcoming words (Kuperberg & Jaeger, 2016). In particular, a number of studies have shown that information at the determiner level generates predictions about the upcoming noun.

Many types of information at the determiner can be used to guide predictions during online speech perception, such as disfluency. Speakers are more likely to be disfluent when

referring to objects with low discourse-dependent probability compared to those with high probability (Arnold & Tanenhaus, 2007), leading to a similar type of probabilistic information as discussed above where discourse-dependent probability is associated with some phonetic outcome (in this case, disfluent outcomes). Arnold and colleagues (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Arnold, Kam, & Tanenhaus, 2007) asked whether this probabilistic information at the determiner influences processing of the upcoming noun, which either has high or low discourse-dependent probability. Listeners in these studies heard the same type of instructions used by Dahan et al. (2002) and related studies, but Arnold et al. included disfluencies in the determiner region (e.g., *Put the grapes below the candle… Now put thee uh candle above the chair*). The results revealed that listeners look more to a discourse-new competitor (in this case, *candy*) than a discourse-given target (here, *candle*) when the determiner was disfluent, indicating that this probabilistic information at the determiner level influences predictive processing. Listeners have also been shown to use disfluency as a predictive cue when associated with other low probability entities (e.g., unfamiliar objects, Arnold, Kam, & Tanenhaus, 2007; low frequency referents, Bosker, Quené, Sanders, & De Jong, 2014).

Fluent phonetic information at the determiner has also been found to guide predictive processing. Salverda, Kleinschmidt, and Tanenhaus (2014) presented listeners with stimuli containing natural coarticulatory information between definite determiners and the following nouns and stimuli with no coarticulatory information at the determiner level. They found that listeners were significantly faster to look to the target when given the coarticulatory information vs. no information. These results reveal that listeners can make predictions based on fine-grained (non-probabilistic) phonetic information free of any disfluency.

Morphology can also serve as a predictive cue in languages with gender-marked determiners, such as French. In an eye-tracking study with French listeners, Dahan, Swingley, Tanenhaus, & Magnuson (2000) presented listeners with instructions to click on objects in a visual display and manipulated whether there was gender marking on the determiner (e.g., plural, gender neutral: *Cliquez sur les boutons* 'click on the buttons' vs. singular, masculine: *Cliquez sur le bouton* 'click on the button'). Dahan et al. found that the gender of a determiner can constrain competition between phonologically-related competitors (e.g., *bouton* 'button' and *bouteille* 'bottle') to such a degree that the competitor with mismatching morphological information (e.g., **le bouteille*) does not actually compete for selection at all; listeners did not look to the competitor significantly more than phonologically-unrelated distractors (e.g., *chien* 'dog'). These results indicate that this morphological information allows listeners to make strong predictions about the identity of the upcoming noun. Other studies have shown similar results for Spanish (Dussias,Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013; Lew-Williams & Fernald, 2010; Grüter, Lew-Williams, & Fernald, 2012) and German listeners (Hopp, 2013, 2015, 2016).

Similar to how variations in discourse-dependent probability influence the rate of disfluency at the determiner (Arnold & Tanenhaus, 2007), discourse-dependent probability also leads to phonetic reduction on fluent determiners that precede high probability content words (Kahn & Arnold, 2012, 2015; Study 1). Given the body of work showing the predictive power of determiners, this type of probabilistic information may signal to the listener that the upcoming target has high or low discourse-dependent probability. One goal of the current study is to investigate this possibility for L1 as well as L2 listeners.

3.1.3 Prediction in L2 Perception

L2 speech perception (like L1 perception) is an incremental process, characterized by competition between multiple lexical candidates prior to lexical selection. As discussed above, listeners rely both on target-concurrent and contextual information to make predictions about the most likely candidate during this competition. Just as the same incremental processes are assumed for L1 and L2 perception, it is generally assumed that the same mechanisms underlie L2 predictive processing as well. However, differences in the degree to which listeners use predictive processing may arise due a number of sources, such as experience with a language (see Kaan, 2014, for a review; see also, Huettig & Mani, 2016; Kuperberg & Jaeger, 2016).

By default, L2 listeners have less experience with a language than L1 listeners of that language. In models of L2 recognition (e.g., BIA; Dijsktra & van Heuven, 1998) this difference in experience has been encoded in lexical representations; the lower resting activation levels for L2 vs. L1 listeners predicts delays in L2 vs. L1 word recognition (e.g., Duyck, Vanderelst, Desmet, & Hartsuiker, 2008). Some studies have suggested that differences like these in L1 vs. L2 lexical representations impact how listeners in each group engage predictive processes during speech perception. For example, Hopp (2013) found that predictive processing ability related to speed of lexical access. Additionally, differences in experience with L2-specific linguistic structures (e.g., grammatical gender) and co-occurrence relations (e.g., semantic relatedness of nouns and verbs) also plays an important role in whether L2 listeners can engage predictive processing. These previous studies provide important insights into the limits of these aspects of predictive processing, and outline the types of differences we might expect to observe when comparing L1 and L2 listeners for other types of predictive processing (e.g., prediction based on

reduction, the focus of the current work). We devote the remainder of this section to the discussion of these studies.

Many studies have shown that L2 listeners struggle with the use of L2-specific morphological information as a predictive cue during speech perception (Dussias et al., 2013; Grüter et al., 2012; Hopp, 2013, 2015, 2016; Lew-Williams & Fernald, 2010; Martin, Thierry, Kuipers, Boutonnet, Foucart, & Costa, 2013). This work has typically investigated whether L2 listeners whose L1 does not have grammatical gender (e.g., English) can use gender marking on determiners to predict the identity of the upcoming noun. In a series of eye-tracking studies with bilingual English and Spanish listeners, participants heard gender-marked determiners in a sentence context. When L1-Spanish listeners encountered a gender-marked determiner (e.g., the feminine determiner *la*), they looked more to an image with a feminine name, indicating that they predicted the upcoming noun was feminine; however, L2-Spanish listeners did not (Dussias et al., 2013; Lew-Williams & Fernald, 2010). Hopp (2013, 2016) has observed similar results for English(L1)-German(L2) bilinguals. Other morphological features drive predictions as well. An ERP (event-related potential) reading study with Spanish-English bilinguals found that L1- but not L2-English listeners used phonological variants of the English indefinite determiner (i.e., *a* vs. *an*) as a predictive cue (Martin et al., 2013).

Critically, studies have shown that L2 listeners can use morphology to make predictions if similar morphological structures exist in the L1. For example, L2-Spanish listeners with grammatical gender in their L1 (e.g., Italian or French) can make predictions based on the gender marking of the determiner (Dussias et al., 2013; Foucart, Martin, Moreno, & Costa, 2014; Foucart, Ruiz-Tada, & Costa, 2015), lending support to the argument that experience influences the ability to engage predictive processing (Kaan, 2014). However, certain conditions must be

met for L2 listeners to make predictions. Dussias et al.'s Italian listeners made predictions based on the feminine, but not masculine, determiner, which could be attributed to differences in the details of the gender systems of Italian and Spanish (specifically, Italian has two masculine determiners but only one feminine). This divergence in the gender systems across languages could have posed a barrier to mastery for those somewhat low proficiency listeners, leading to cross-language competition between determiner systems (Morales, Paolieri, Dussias, Valdés Kroff, Gerfen, & Bajo, 2016). Similarly, mastery of the L2 (German) determiner system was critical for listeners in Hopp (2013, 2016), who lacked grammatical gender in their L1 (English). Listeners with strong mastery (assessed via a production task) showed similar levels of prediction as L1 listeners, while listeners with weak mastery did not predict. Furthermore, while listeners with general mastery of the language (i.e., high proficiency listeners) are able to make predictions (Dussias et al., 2013), they make less consistent predictions than L1 listeners (Grüter et al., 2012). Together, these results suggest that when L2 listeners must make predictions based on language-specific knowledge, they are often unable to do so as well as L1 listeners, in some cases even when their own L1 requires similar knowledge.

Semantic information also seems to influence L1 and L2 predictive processing in distinct ways. In an eye-tracking study, Dijkgraff, Hartsuiker, and Duyck (2016) presented listeners with sentences that either had a strong relationship between the verb and target noun (e.g., *reads* and *letter*) or a neutral relationship (e.g., *steals* and *letter*). Previous studies have found that L1 listeners look more to the target earlier in processing (before target onset) in strong vs. neutral conditions, suggesting that listeners predict the identity of the noun after hearing the verb (Altmann & Kamide, 1999). Dijkgraff et al. replicated these results with L1 listeners, and also found that L2 listeners (L1-Dutch) showed the same overall bias toward the context-appropriate

nouns. However, the groups differed in the time-course of this effect; L2 listeners were slower to predict the appropriate noun than L1 listeners. These results suggest that L2 listeners can engage predictive processing but they may differ from L1 listeners in how predictive processing proceeds over time, perhaps due to slower or weaker lexical access for L2 vs. L1 listeners (Kaan, 2014; Shook, Goldrick, Engstler, & Marian, 2014). Consistent with this possibility that strong, highly automatic lexical access is necessary for L1-like predictive processing, Hopp (2013) observed that listeners with fast response times in a baseline condition (indexing the overall speed of lexical access) showed more rapid onset of predictive effects in these critical conditions.

However, results from Shook et al. (2014) suggest that L2 listeners can make use of global semantic context in a similar manner as L1 listeners, despite having weaker lexical access overall. Listeners in this study identified target words in predictive sentences, which created a context where the target word had high probability (e.g., *a can* has high probability in *The drinker went to the recycling bin and threw away the bottle and…*). Eye-movements to the target in this condition were compared to those when the target had low probability in non-predictive contexts (e.g., *a can* in *The typist went to the new conference room and brought along a printer and…*). In non-predictive contexts, L2 listeners activated target items more slowly than L1 listeners. However, Shook et al. found that this difference in the strength of lexical access across L1 and L2 listeners was eliminated when the target had high probability in the sentence context, indicating L2 listeners engaged predictive processes in high vs. low probability contexts (see also Chambers & Cooke, 2009).

In contrast to research with L1 listeners, to our knowledge no research has considered whether L2 listeners can use (target-concurrent) discourse-dependent probabilistic reduction for predictive processing. It is, therefore, unclear whether L2 listeners can use this type of

information at all, either at the determiner or noun level. However, there is evidence that L2 speakers produce similar levels of discourse-dependent probabilistic reduction on nouns as L1 speakers (Baker et al., 2011; Study 1), but do not reduce determiners according to their discourse-dependent probability (Study 1). Pickering and Garrod (2013) have argued that there are strong links between the production and perception systems, and that prediction in perception is, at least partially, driven by aspects of the language production system. Under this account, we expect that L2 listeners should be able to make predictions based on the same types of probabilistic information they have no issues producing (consistent with Hopp, 2013, 2016) but have challenges utilizing information they do not utilize in production.

3.1.4 The Current Study

The current study investigates (1) whether L2 listeners differ from L1 listeners in how probabilistic information contributes to predictive processing of concurrent information during online speech perception, and (2) whether listeners use discourse-dependent probabilistic reduction at the determiner level to make predictions about the upcoming noun. To address these two goals, we use a visual world eye-tracking experiment, in which listeners follow a series of instructions to move objects in the visual display (e.g., *Put the candle below the square… Now put the candle above the diamond*). We compare the overall rate of fixations to the target (e.g., *candle*) and its phonologically-related competitor (e.g., *candy*), as well as how fixations change over time, across discourse-new (the target word is present only in the second instruction) and discourse-given (the target word is present in both instructions, as above) trials. Targets were either reduced or unreduced, creating either a congruent or incongruent coupling of discourse-dependent probability and reduction in each trial. Differences across groups in the overall rate of fixations to either the target or competitor, or in the time-course of fixations, speak to the first

goal of the study, allowing us to determine whether L1 and L2 listeners differ in how probabilistic information influences speech perception. Differences in fixations early in processing (prior to the target onset) speak to the second goal of the study, and will indicate that listeners use reduction at the determiner to predict the identity of the upcoming noun. (See Table 3.4 below for a summary of these hypotheses and predictions.)

3.2 Experiment 1

3.2.1 Method

3.2.1.1 Participants. A total of 64 listeners participated in this experiment. All participants also took part in a companion speech production study that used the same picture stimuli. The production study was conducted no less than one week prior to the current study to reduce the influence of recent exposure to the target items. Because hearing examples of phonetic variation in particular conditions could influence subsequent productions, all participants completed the production study first. Participants were divided into two groups: native listeners of American English (henceforth, the L1 group) and native listeners of Mandarin who learned English as a second language (henceforth the L2 group).

Listeners comprising the L1 group (N = 25) were recruited from the linguistics participant pool at Northwestern University and received partial course credit. All participants in this group were L1 listeners of English with no history of speech impairments or color blindness. One participant was excluded due to poor equipment performance, leaving 24 native listeners of American English (17 female; mean age: 19, range: 18-21) for the L1 group.

Many listeners included in the L2 group were recruited from the International Summer Institute at Northwestern University, a month-long program for incoming international students that offers intensive English instruction and one-on-one tutoring prior to the start of their first

academic quarter as graduate students. Other participants were recruited from the Northwestern community via flyers and a database of current and former students in Northwestern's English Language Learners Program. Each of these participants was compensated $10/hour. Two participants were recruited via the linguistics department participant pool, and thus received partial course credit for their participation.

Twenty-eight listeners whose L1 was Mandarin participated as part of the L2 group. Three participants were excluded due to poor equipment performance, and one English-dominant listener was excluded as well. The remaining 24 participants (18 female; mean age: 23.2, range: 18-31) were Mandarin-dominant, L2 listeners of English who did not learn English at home (i.e., English exposure began at school).

Each participant completed a detailed language background questionnaire. In this questionnaire, participants were asked to provide information about their exposure and experience with all languages they spoke. Table 3.1 reports information provided by participants that summarizes variation in linguistic experience across groups. The LexTale vocabulary test (an unspeeded lexical decision task; Lemhöfer & Broersma, 2012) was used as an objective measure of English proficiency. Additional details about the LexTale task are reported below.

Table 3.1. Language background information. Mean (standard deviation).

	Percent Correct, LexTale vocabulary test	Age of first exposure to English (years)	Length studied English (years)	Percent time English used	Self-rated speaking ability (Perfect:10)	Self-rated listening ability (Perfect:10)
L1 group (N = 24)	95.7 (5.2)	0 (0)	18.9 (1.2)	95.7 (5.8)	9.7 (0.6)	9.8 (0.5)
L2 group (N = 24)	80 (13.4)	7.9 (5.8)	15.6 (6.8)	48.6 (27.4)	5.8 (2.4)	6.6 (2.3)

3.2.1.2 Materials and design. The stimuli were a set of 48 pictures taken from the Bank of Standardized Stimuli (BOSS; Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010) and other sources as needed. The BOSS database includes a large set of full color photographs that have been normed along a number of dimensions, including name agreement, category, familiarity, and visual complexity. The names for the target pictures have high name agreement by L1 speakers (mean = 94.6%, min = 63.2%, sd = 9.2%), as established in a separate norming study with 19 participants. The names also have above at least 40% name agreement by L2 speakers from the same population as the participants for the current study, also established in a separate norming study with 10 participants (mean = 77.9%, sd = 20.3%).

The stimuli were put into 24 pairs that minimally share the first two phonemes (e.g., /kæn-/ in *candy* and *candle*) and have the same number of syllables. The items in each pair were assigned the role of target or competitor such that the mean lexical frequencies (according to the SUBTLEX-US corpus; Brysbaert & New, 2009) of targets and competitors were not significantly different according to a t-test (targets: mean = 54.9 occurrences per million, sd = 104.7; competitors: mean = 35.9 occurrences per million, sd = 34.5; $t(27.951) = 0.847$, $p > 0.05$). Each pair of items was assigned two phonologically and semantically unrelated distractors (also picturable nouns), which appeared in the display with the target and competitor items. See Appendix A for the full set of stimuli, with name agreement and frequency data for each target item.

Two variables were manipulated in the stimuli: discourse status of the target (discourse-given based on the first instruction, or discourse-new after the first instruction), and reduction of the target noun and its determiner (reduced or unreduced). Experimental trials included instructions to move the target item above or below one of the geometric shapes located in each

corner of the display. Listeners heard two instructions per trial, and the second instruction included the critical region for analysis. The experiment included 24 target trials, one trial for each pair of target items. Half of the target trials were assigned to the discourse-given condition and half to the discourse-new condition. Half of the trials in each of these conditions were assigned to the reduced condition, and the other half to the unreduced condition. Four lists were created to vary the four experimental conditions in which each item was presented, and an additional four lists were created to reverse the roles of the target and competitor items (i.e., *candy* as target and *candle* as competitor in one list, but *candle* as target and *candy* as competitor in another). Three participants were randomly assigned to each list, and each participant received a randomly generated order for that list.

In addition to the 24 target trials, listeners heard 28 filler trials, which were constant across the eight experiment lists. Twelve of the filler trials included two phonologically related and two unrelated items (e.g., *backpack, balloon, clock, finger*). Of these 12 trials, six included one of the phonologically related items in the first instruction and one of the unrelated items in the second instruction. The other half had unrelated items in both instructions. For the other 16 filler trials, all four items were phonologically unrelated (e.g., *anchor, match, ball, radio*). Half of the 28 filler trials were assigned to the discourse-given condition, the other half to the discourse-new condition. Four of the filler trials were used as practice, with two each for each condition.

3.2.1.3 Recordings. The instructions were produced by a female, L1 speaker with an Inland North American accent. Recordings were made in a sound-attenuated booth using a boom-mounted Shure SM81 Condenser Handheld Microphone sampling at 44,100 Hz and SoundStudio software. Following previous work (e.g., Dahan et al., 2002), the productions were

elicited by reading from printed instructions. For the unreduced conditions, the target determiner-noun sequence was printed in capital letters. Three sets of sentences were produced for each trial. Each set of sentences included two instructions, separated by a semicolon to elicit a rising continuation intonation. One type of sentence included a repetition of the target item, to felicitously elicit discourse-dependent reduction (e.g., *Put the candy below the triangle; Now put the candy above the square*). Another type of sentence included the competitor item in the first instruction and one of the unrelated items in the second instruction, to establish a discourse-new context (e.g., *Put the candle below the triangle; Now put THE BOOK above the square*). The final sentence type included the target sequence in capital letters in the second instruction and an unrelated item in the first instruction, to elicit an unreduced production without a contrastive production that could occur with the related competitor in the first instruction (e.g., *Put the book below the triangle; Now put THE CANDY above the square*). The underlined sentences were excised and combined in various configurations to create the trial instructions.

Each utterance was measured to ensure there were substantial differences in duration between the target determiners and nouns in reduced vs. unreduced conditions, which was critical for the reduction manipulation. Furthermore, the duration of the words in the second instruction preceding the target sequence (i.e., *Now put*) as well as the pauses between these words and the target sequence were measured. These durations were examined to determine whether there were differences in the acoustic information available to participants prior to the target sequences in the reduced vs. unreduced conditions. Results of these measurements are shown in Tables 3.2 and 3.3.

Table 3.2. Mean durations of target noun and determiner in each condition (ms). Standard deviation in parentheses.

	Reduced condition	Unreduced condition
Determiner	47.9 (11.7)	181.5 (47.5)
Target noun	299.2 (62.7)	535.9 (102.3)

Table 3.3. Mean durations of words and pauses in preamble (ms). Standard deviation in parentheses.

	Reduced condition	Unreduced condition
now	132.8 (30.4)	124.8 (29.1)
Pause following *now*	60.1 (8.0)	65.8 (8.6)
put	71.8 (13.9)	145.3 (26.3)
Pause following *put*	56.9 (10.5)	111.7 (71.6)

The mean duration of determiners in the reduced condition (47.9 ms) was 133.6 ms shorter than the mean of those in the unreduced condition (181.5 ms). To determine whether this difference was reliable, a linear mixed effects regression was built with determiner duration as the dependent variable, reduction condition as a contrast-coded fixed effect, and a random intercept for item (i.e., the target noun). The effect of reduction condition on determiner duration was significant according to a model comparison ($\beta = -0.134$, se = 0.007, $\chi^2(1) = 150.85$, $p < 0.001$). Similarly, there was a substantial difference in the mean duration of target nouns in the reduced vs. unreduced condition (difference: 236.7 ms; reduced: 299.2 ms; unreduced: 535.9 ms). A similar linear mixed effects regression was built with noun duration as the dependent

variable. The effect of reduction condition on noun duration was significant according to model comparison ($\beta = -0.237$, se = 0.01, $\chi^2(1) = 132.57$, $p < 0.001$). These results confirm that listeners were exposed to condition-appropriate phonetic variation.

It was also important to examine a series of durations preceding the critical region of the instruction, namely, the duration of *now* and *put*, and the duration of the pause between *now* and *put* as well as the pause between *put* and *the*. Any condition-dependent differences in any of these areas could provide listeners with cues to reduction condition prior to the critical region, which would compromise our results. There was minimal difference in the duration of *now* in the reduced vs. unreduced condition (difference: 3.9 ms; reduced: 124.8 ms; unreduced: 132.8 ms). A linear mixed effects regression with the duration of *now* as the dependent variable, reduction condition as a contrast-coded fixed effect, and a random intercept for item (i.e., target noun) confirmed there was no significant difference in *now* duration ($\beta = -0.008$, se = 0.006, $\chi^2(1) = 1.75$, $p > 0.05$). There was also a small difference in the duration of the pause between *now* and *put* across reduction condition (difference: 5.7 ms; reduced: 60.1 ms; unreduced: 65.8 ms). However, a similar linear mixed effects regression with the duration of the pause as the dependent variable revealed that this difference was significant ($\beta = -0.006$, se = 0.002, $\chi^2(1) = 11.03$, $p < 0.001$). The mean duration of *put* in the reduced condition (71.8 ms) was 73.5 ms shorter than that in the unreduced condition (145.3 ms). A similar linear mixed effects regression with *put* duration as the dependent variable indicated a significant difference ($\beta = -0.073$, se = 0.004, $\chi^2(1) = 135.73$, $p < 0.001$). Finally, the mean duration of the pause between *put* and *the* was substantially shorter in the reduced vs. unreduced condition (difference: 54.8 ms; reduced: 56.9 ms; unreduced: 111.7 ms). A similar linear mixed effects regression with the duration of the

pause as the dependent variable confirmed that this difference was significant ($\beta = -0.055$, se = 0.01, $\chi^2(1) = 24.69$, $p < 0.001$).

These analyses revealed that there were significant acoustic differences in the region directly preceding the critical region for analysis (the preamble). To ensure that any effects of the reduction manipulation on eye movements were due to variation in the critical region, the preamble from the reduced condition was cross-spliced with the critical region from the unreduced condition, and vice versa. Prior to cross-splicing, the intensity of all sound files (including filler trials, which were not cross-spliced) were normalized to 66 dB. The pre-critical regions were then extracted and the intensity between the end of *put* and the beginning of *the* was leveled to eliminate voicing leading into a number of unreduced determiners. Then, these regions were spliced in to the appropriate file. After cross-splicing, all of the files were again normalized for intensity to 66 dB. Following this procedure, the first author verified that the file did not sound extremely unnatural. However, cross-splicing may have resulted in disruptions to overall prosody that could impact listeners' perception behavior (Dilley & McAuley, 2008). We return to this point in the discussion below.

3.2.1.4 Procedure. Participants were seated in front of a computer screen while their eye movements were recorded with an SR Research EyeLink 1000 plus eye-tracker sampling at 1 kHz. Participants received auditory input over Sony MDR-7506 headphones, with the instructions played at a volume comfortable to the participants. Presentation of the audio and visual components of the experiment, as well as data collection, were controlled by the SR Research Experiment Builder program.

Participants were presented with written instructions, and were given as much time as needed to read them. The experimenter confirmed that all instructions were understood, and

confirmed that participants in the L2 group were familiar with the names of the geometric shapes. After the instructions, the eye-tracker was calibrated. The experiment began with four practice trials (described above), followed by a randomized set of 24 target and 24 filler trials. Prior to each trial, a circle appeared in the center of the screen to correct any drift in participants' eye position since the initial calibration. Participants were instructed to fixate on the circle and press the spacebar on the keyboard in front of them to start the trial. At the onset of each trial, a 5x5 grid appeared with four colored images (a target, a cohort competitor, and the two unrelated distractors; e.g., *brick, bridge, airplane, hand*) arranged in the center (in an order determined in the experiment list, and counterbalanced across lists) and four geometric shapes (circle, square, triangle, and diamond) arranged at the edges of the display. A fixation cross was positioned in the center of the grid, surrounded by the colored images (Figure 3.1). After 500 ms, the first spoken instruction (i.e., the preamble) began (e.g., *Put the brick below the triangle*). The participant was instructed to complete the action by using a mouse to moving the picture of the candy to the position in the grid below the triangle. As soon as the participant completed the action, the second spoken instruction began after 500 ms (e.g., *Now put the brick above the square*). After completing the second action, the drift correct circle appeared and the next trial began. The experiment took roughly 10 minutes.

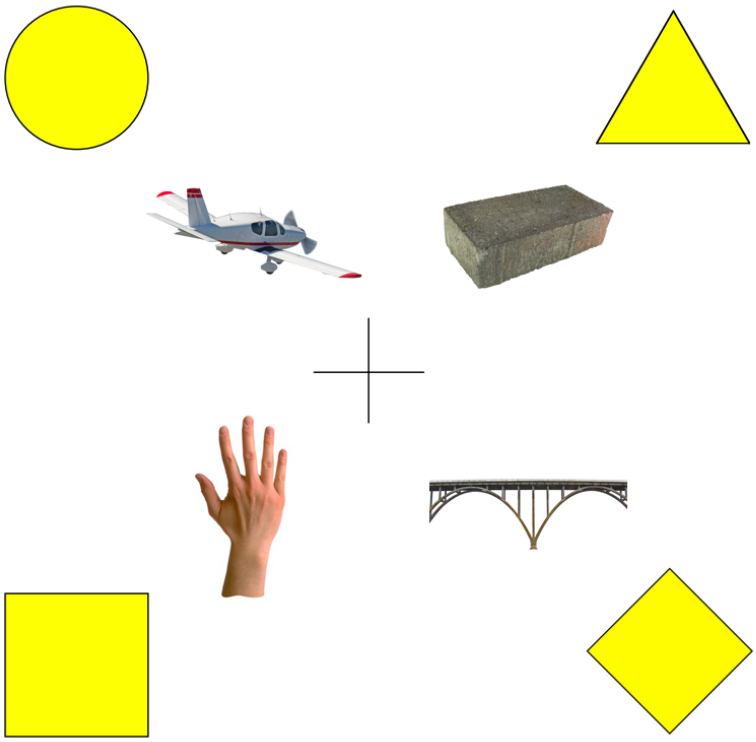

Figure 3.1. Example visual display from experiment.[6]

3.2.1.5 Additional Tasks. Following the experiment, participants were shown sets of four pictures (a target, cohort competitor, and the two unrelated distractors from each target trial) that they had just seen in the experiment and were asked to choose the picture that depicted a word printed at the top of the screen. Each set of four pictures was presented twice to test participants' knowledge of both the target and the cohort competitor, as it was critical for participants to know both of these words to interpret the eye tracking results.

Additionally, participants completed the LexTale receptive vocabulary test (Lemhöfer & Broersma, 2012), which has been independently tested and validated as an objective measure of

[6] Images are from the Bank of Standardized Stimuli (BOSS; Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010) and are authorized for redistribution according to the Creative Commons Attribution-Share Alike 3.0 license (https://creativecommons.org/licenses/by-sa/3.0/).

English profiency. The LexTale test was presented in Praat. Participants read the instructions, which indicated that they should evaluate whether a string of characters was an existing word of English. They were instructed to choose 'yes' when they were sure the string was an existing word, even if they were not sure of the meaning of the word. They were instructed to choose 'no' if they were not sure the string was an existing word. Unlike typical lexical decision tasks, the instructions indicated that participants could take as much time as they needed to make their decision. Given this instruction, response times to lexical decisions were not considered. Instead, an average accuracy measure was considered ((number of words correct/total number of words * 100) + (number of nonwords correct / total number of nonwords * 100) / 2). Descriptive statistics for the LexTale results are summarized in Table 3.1.

3.2.1.6 Data pre-processing. A series of pre-processing steps were performed to prepare the data for analysis. Analysis of eye movements was only performed on accurate trials. Trials were excluded if listeners made recognition errors in either phase of trials (failed to select the correct picture), moved pictures to incorrect positions in either phase of trials, or made errors matching the target or competitor with the corresponding name in the post-test. In total, 121 trials (10.5%) were excluded from the eye movement analysis under these criteria. These trials were included in the accuracy analysis below.

For the eye movement analysis, additional trials were excluded when the equipment failed to record eye movements for large proportions of the trial. Individual trials were excluded if more than 25% of samples failed to track the eye appropriately (as assessed by the EyeLink software). Furthermore, entire participants were excluded if more than 40% of their trials had been excluded due to the preceding criterion (N = 1). Finally, individual trials in which participant did not look at the target within the critical window of analysis (i.e., 0-1500 ms form

the onset of target noun production) were excluded. Following this final exclusion of individual trials, participants who only had one observation in any given experimental condition were excluded from analysis (N = 2), as it would be impossible for statistical models to accurately estimate variance in a condition with a single observation in a cell. All participants excluded for these reasons were replaced with new participants from the same population. Ultimately, one L1 and two L2 participants were replaced, and 52 trials (4.5%) were excluded from analysis due to trackloss.

Finally, one target-competitor pair was removed from both accuracy and eye movement analyses due to an error in stimulus design (*scar* and *star* are rhyme competitors, not cohort competitors). This final exclusion resulted in removal of 48 trials (4.3%). Following these exclusions, 931 trials (80.8%) remained for inclusion in the eye movement analysis.

3.2.1.7 Accuracy analysis. The rate of errors made in the post-test were analyzed using a mixed-effects logistic regression, with a binary (correct vs. incorrect) dependent variable and a contrast-coded fixed effect for group. The maximal random effects structure justified by the data (Barr, Levy, Scheepers, & Tily, 2013), including a decorrelated by-item random slope for group, and a random intercept for participant.

The rate of recognition errors produced by L2 participants (L1 listeners made no recognition errors) in the second (critical) phase of the trial was considered in an analysis using logistic regression. The dependent variable in this model was a binary measure (correct vs. incorrect). Fixed effects included contrast-coded effects for discourse-condition (discourse-given vs. discourse-new) and reduction condition (reduced vs. unreduced), as well as their interactions. Nested model comparison was used to perform significance tests for all accuracy analyses.

3.2.1.8 Eye movement analysis. Four models were built to analyze looks to the target and competitor in an early vs. late window (0-199 ms and 200-1500 ms following the onset of noun production, respectively). As in previous work (e.g., Allopenna et al., 1999; Dahan et al., 2002), we assume that fixations prior to 200 ms after stimulus onset were driven by speech preceding that onset. Therefore, we assume that looks in the early window were driven by responses to the determiner, while looks in the late window were driven by the target. Fixations to the target and competitor in the early window were averaged across the entire window, while fixations in the late window were collapsed across 50 ms time bins. For all analyses, linear mixed effects regressions were built using unweighted empirical logits of fixations within these time bins.[7] Fixation proportions were transformed to empirical logits in order to correct for the bounds on proportions that are problematic for logistic regression (for further discussion, see Barr, 2008).

Models for analyzing target and competitor looks in the early window included contrast-coded fixed effects for group (L1 vs. L2), discourse condition (discourse-given vs. discourse-new) and reduction condition (reduced vs. unreduced). Both models included the maximal random effects structure supported by the data (Barr et al., 2013). For the target model this included decorrelated random slopes for discourse condition, reduction condition, and their interaction for the by-participant intercept. The competitor model included only a random intercept for participant, with no random slopes.

[7] Alternative growth-curve analysis methods were rejected due to poor performance. Weighted empirical logit models yielded fits that were wildly divergent from the observed responses. Analyses with logistic regression would only converge with random intercepts, and such models are far too anti-conservative for this type of analysis.

Eye-movements to both the target and competitor in the late window were analyzed using growth curve analysis (GCA), which utilizes linear mixed effects regressions to model changes in eye movement over time (Mirman, 2014). This analysis technique captures non-linear changes in behavior over time by considering the higher-order polynomial terms of a curve. Furthermore, this technique involves creating orthogonal polynomial terms, allowing for independent evaluation of each term that would otherwise be dependent on each of the other terms. Each polynomial term is then entered as a fixed effect in the regression as a predictor of changes in eye movement over time.

Both models for target and competitor looks in the late window included polynomial terms up to a quartic term. In order to determine the highest order polynomial term to include in these models, we began with the lowest order term and added additional terms until we achieved reasonable model fit (assessed by plotting fits against observed data). Contrast-coded fixed effects for discourse condition (discourse-given vs. discourse-new), reduction condition (reduced vs. unreduced), and group (L1 vs. L2) were also included as predictors, as well as their interactions with each other and each polynomial term. Models included the maximal random effects structure supported by the data, with random intercepts for participant (random slopes for items were not possible, as the empirical logit transformation requires collapsing across items) and decorrelated by-participant random slopes for each polynomial term. Significance of main effects and interactions for all models in both windows was assessed via nested model comparison.

3.2.1.9 Eye movement hypotheses and predictions. Table 3.4 summarizes the main hypotheses for the current study, with corresponding predictions and expected results. The critical prediction effect expected based on previous work (e.g., Dahan et al., 2002) is facilitation

of perception by the appropriate coupling of discourse-dependent probability and reduction. When targets have high discourse-dependent probability, congruent coupling constitutes a reduced production; targets with low discourse-dependent probability are congruently coupled with an unreduced production. Because the appropriate level of reduction shifts as a function of discourse-dependent probability, we therefore predict an interaction of these two factors. Results in both the early and late time windows can be understood in terms of these hypotheses. However, predictions relating to time course differences do not apply to results in the early window.

Table 3.4. Hypotheses and predictions for eye movement analyses.

Hypotheses	Predictions	Expected results	Model output
Difficulties in L2 speech perception and/or differences in L1 and L2 experience create difficulties in L2 predictive processing	Differences across groups in the prediction effect: the facilitation of perception by congruent coupling of discourse-dependent probability and reduction	Overall magnitude of prediction effect differs across groups	Interaction of group by discourse condition by reduction condition
		Prediction effect unfolds differently over time across groups	Interaction of group by discourse condition by reduction condition at polynomial time terms
L2 listeners have no difficulty processing reduction in the L2 and/or L1 and L2 listeners have substantially similar experience with discourse-dependent reduction	No differences across groups in prediction effect	No difference across groups in overall magnitude of prediction effect	No interaction of group by discourse condition by reduction condition
		No difference across groups in how prediction effect unfolds over time	No interaction of group by discourse condition by reduction condition at any polynomial time term

3.2.2 Results

3.2.2.1 Accuracy. L1 listeners made five errors during the post-test, amounting to a 0.4% error rate, while L2 listeners made 16 errors (1.4%). Results from the mixed effects logistic regressions revealed no main effect of group ($\beta = 0.09$, SE = 1.47, $\chi^2(1) = 0.002$, $p > 0.05$). This indicates that listeners across group were equally familiar with the stimuli included in the experiment, and ensures that any group differences in subsequent error and eye movement analyses cannot be due familiarity with the stimuli.

Table 3.5 shows recognition error rates for the second (critical) phase trials for L2 listeners, only for trials in which no errors were made in the first phase of the trial and no errors were made for the target and competitor items in the post-test. There was no significant difference in the proportion of errors made in trials with a reduced vs. unreduced target ($\beta = 6.91$, SE = 774.6, $\chi^2(1) = 0.27$, $p > 0.05$). Listeners made marginally more errors in discourse-new vs. discourse-given trials ($\beta = 8.75$, SE = 774.6, $\chi^2(1) = 3.53$, $p < 0.07$). However, this effect was modulated by a significant interaction between givenness condition by reduction condition ($\beta = 19.02$, SE = 1549.11, $\chi^2(1) = 8.36$, $p < 0.01$). Follow up regressions revealed that in discourse-new trials, listeners mis-recognized the target more it was reduced vs. unreduced ($\beta = -2.60$, SE = 1.04, $\chi^2(1) = 11.79$, $p < 0.001$). In discourse-given trials, a marginal effect showed the opposite pattern, where listeners made somewhat more errors when the target was unreduced vs. reduced ($\beta = 17.42$, SE = 2254.05, $\chi^2(1) = 2.82$, $p < 0.1$).

Table 3.5. Mean recognition error rates for L2 listeners across conditions and phases of the trial in Experiment 1. Standard error in parentheses.

	Discourse-given	Discourse-new
Reduced	100% (0%)	90.9% (1.9%)
Unreduced	98.1% (1.3%)	99.3% (0.7%)

Interim Summary. In all cases, errors were mis-selections of the competitor picture. Altogether, these error results indicate that L2 listeners were sensitive to the coupling of the discourse status and reduction of the target. Listeners were especially likely to mis-select the competitor picture when the target was reduced in discourse-new conditions. This suggests that the presence of reduction is a strong cue to L2 listeners that the target is likely to be discourse-given, leading to a failure to reject incorrect predictions and to selection of the competitor (which was discourse-given after the first phase of the trial). A similar, although less reliable, effect was seen for the discourse-given condition, where listeners chose the competitor in error somewhat more often when the target was unreduced vs. reduced.

3.2.2.2 Eye movements in the early window. Overall proportions of fixations to the target and competitor in the early window are shown in Tables 3.6 and 3.7. Neither group of listeners showed facilitation of perception when discourse-dependent probability and reduction were appropriately coupled within the context preceding the target. While listeners showed an overall bias to look at discourse-new images in the display (for both the target and the competitor), they showed no sensitivity to the level of reduction on the target.

This pattern of results was substantiated by the regression analysis. A main effect of discourse condition on looks to both the target and competitor indicated that listeners looked significantly more to the target in discourse-new vs. discourse-given conditions ($\beta = -0.76$, SE = 0.15, $\chi^2(1) = 20.29$, $p < 0.001$), and more to the competitor in discourse-given vs. discourse-new conditions ($\beta = 0.51$, SE = 0.13, $\chi^2(1) = 14.62$, $p < 0.001$). For looks to the competitor, this main effect was modulated by an interaction with group ($\beta = 0.58$, SE = 0.26, $\chi^2(1) = 4.85$, $p < 0.05$), indicating that the overall bias to look at discourse-given competitors was much stronger for L1

($\beta = 0.81$, SE = 0.18, $\chi^2(1) = 18.86$, $p < 0.001$) relative to L2 listeners ($\beta = 0.23$, SE = 0.19, $\chi^2(1) = 1.43$, $p > 0.05$). No other fixed effects or interactions were significant for looks to the target or competitor (all $\chi^2(1) > 1.8$, ps > 0.05).

Table 3.6. Mean proportion of looks to the target in the early time window (0-199 ms) across discourse conditions, reduction conditions, and groups in Experiment 1. Standard error in parentheses.

	Discourse-given		Discourse-new	
	Reduced	Unreduced	Reduced	Unreduced
L1	8.1% (2.6%)	9.7% (3.1%)	28.0% (4.1%)	26.7% (3.1%)
L2	10.9 (2.8%)	8.9% (2.8%)	22.3% (3.7%)	25.0 (3.5%)

Table 3.7. Mean proportion of looks to the competitor in the early time window (0-199 ms) across discourse conditions, reduction conditions, and groups in Experiment 1. Standard error in parentheses.

	Discourse-given		Discourse-new	
	Reduced	Unreduced	Reduced	Unreduced
L1	23.3% (4.0%)	27.1% (4.2%)	10.5% (2.9%)	7.1% (2.3%)
L2	25.2% (3.9%)	29.7% (4.6%)	9.0% (2.9%)	15.3 (4.1%)

3.2.2.3 Eye movements to the target in the late window.

3.2.2.3.1 ***Fixations overall.*** Before considering the shapes of fixation curves, it is useful to consider influences of the experimental manipulations on the overall proportion of looks to the target and competitor (Figure 3.2). Note that for this and all subsequent sections, we begin with an intuitive overview of the results that summarizes the outcome of our regression modeling. We then turn to the details of the growth curve modeling analyses that provide rigorous quantitative support for these summaries.

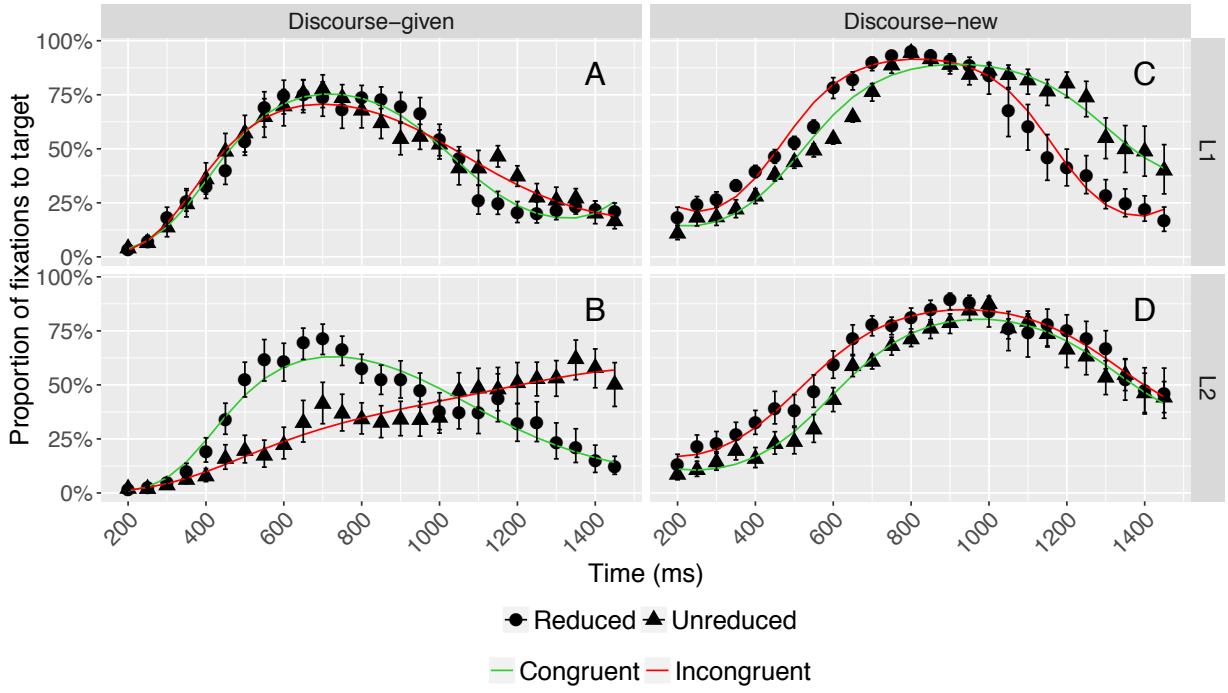

Figure 3.2. Proportion of looks to the target in the late window across groups (horizontal), discourse conditions (vertical), and reduction conditions (shape) in Experiment 1. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-new). Error bars show standard error.

Listeners from each group fixated to the target for similar proportions of time; there was no substantial difference in the overall proportion of looks to the target in Figure 3.2A/3.2C vs. Figure 3.2B/3.2D. Collapsing across groups, listeners tended to look to the target more when it was reduced vs. unreduced (circles above triangles), but this effect was much larger for L2 listeners (for example, proportion of looks to reduced target far exceeds looks to unreduced target in Figure 3.2B). Furthermore, listeners looked substantially more to the target when it was discourse-new vs. discourse-given, but the difference across conditions was larger for L2 vs. L1 listeners (i.e., the overall proportion of looks to the target was higher in Figure 3.2C vs. 3.2A and for 3.2D vs. 3.2B, but more so for 3.2D vs. 3.2B). However, there were no substantial prediction effects. In terms of overall looks to the target, both groups responded similarly to the coupling of

discourse-dependent probability and reduction (but see the discussion of time-course effects below).

Growth curve analysis. These observations were supported by the regression analysis. There was no main effect of group ($\beta = 0.36$, SE = 0.24, $\chi^2(1) = 2.07$, $p > 0.05$), indicating no difference in the overall amount that L1 and L2 listeners fixated on the target (Figure 3.2A/3.2C vs. 3.2B/3.2D). However, there were significant main effects of reduction condition ($\beta = 0.12$, SE = 0.06, $\chi^2(1) = 4.43$, $p < 0.05$) and givenness condition ($\beta = -0.99$, SE = 0.06, $\chi^2(1) = 311.99$, $p < 0.001$), indicating that listeners looked more overall to the target when reduced vs. unreduced (circles vs. triangles) and when discourse-new vs. discourse-given (Figure 3.2C/3.2D vs. Figure 3.2A/3.2B).

The simple main effects of discourse condition and reduction condition were modulated by interactions with group (reduction by group: $\beta = -0.41$, SE = 0.11, $\chi^2(1) = 12.81$, $p < 0.001$; discourse by group: $\beta = 0.28$, SE = 0.11, $\chi^2(1) = 6.53$, $p < 0.05$). Follow-up regressions revealed that L2 but not L1 speakers look significantly more at the target when reduced vs. unreduced (L1: $\beta = -0.09$, SE = 0.08, $\chi^2(1) = 1.42$, $p > 0.05$; L2: $\beta = 0.32$, SE = 0.08, $\chi^2(1) = 15.65$, $p < 0.001$; circles above triangles in Figure 3.2B/3.2D but not Figure 3.2A/3.2C). Furthermore, while both L1 and L2 listeners looked significantly more at the target in discourse-new vs. discourse-given conditions, the effect was larger for L2 listeners (L1: $\beta = -0.85$, SE = 0.08, $\chi^2(1) = 125.80$, $p < 0.001$; L2: $\beta = -1.14$, SE = 0.08, $\chi^2(1) = 187.31$, $p < 0.001$; difference in proportion larger in Figure 3.2D vs. 3.2B than Figure 3.2C vs. 3.2A). However, there was no discourse condition by reduction condition interaction ($\beta = -0.07$, SE = 0.11, $\chi^2(1) = 0.38$, $p > 0.05$). Furthermore, the three-way interaction with group also failed to reach significance ($\beta = 0.30$, SE = 0.22, $\chi^2(1) = 1.88$, $p > 0.05$).

Interim Summary. Overall, neither group exhibited effects consistent with the hypothesis that listeners use discourse-dependent probabilistic reduction to make predictions during speech perception. That is, listeners did look significantly more to the target when presented with congruent probabilistic information (e.g., discourse-given and reduced) vs. incongruent information (e.g., discourse-given and unreduced). The expected prediction effect (a discourse condition by reduction condition interaction) was not observed, nor was there a three-way interaction with group that would indicate differences in how groups engage predictive processing (see Table 3.4). The next analysis considers whether the expected prediction effects can be observed when considering how looks to the target change over time, and whether groups exhibit differences in the time course of prediction effects.

3.2.2.3.2 ***Fixations over time: Rate of looking towards and away from the target.*** Next we consider how changes in the shape of the fixation curves over time may differ across experimental conditions and/or groups. We first focus on factors influencing the rate at which listeners look to the target and how quickly they begin looking away. There was a substantial difference across groups how quickly listeners looked to the target (steeper lines in Figure 3.2A/3.2C vs. 3.2B/3.2D). Furthermore, while both groups of listeners looked more quickly to the target when it was discourse-new vs. discourse-given, the difference across conditions was especially large for the L2 vs. L1 listeners (steeper lines in Figure 3.2C vs. 3.2A, more so in Figure 3.2D vs. 3.2B). Collapsing across groups, listeners looked more quickly to the target and then began looking away from it more quickly when it was reduced vs. unreduced (steeper rise and sharper peak for circles vs. triangles). Furthermore, listeners began looking away from the target more quickly when there was a congruent vs. incongruent coupling of discourse-dependent probability and reduction in discourse-given trials (sharper peak for green vs. red lines in Figure

3.2A/3.2B). A similar, although less pronounced, difference across reduction conditions can be seen for discourse-new trials as well (sharper peak for green vs. red lines in Figure 3.2C/3.2D).

Critically, predictive effects (enhanced perceptual processing with congruent coupling of reduction and probability) differed across groups; only the L2 group exhibited the expected sensitivity. When the target had high discourse-dependent probability (discourse-given conditions), L2 listeners looked more quickly to the target when it was reduced vs. unreduced (steeper slope for green vs. red line in Figure 3.2B). Furthermore, listeners looked away from the target more quickly when the coupling was congruent vs. incongruent (sharper peak for green vs. red line in Figure 3.2B). L1 listeners showed no such effects (no substantial differences in red and green lines in Figure 3.2A). Neither group showed evidence of predictive effects when targets had low-discourse dependent probability (Figures 3.2C and 3.2D).

Growth curve analysis. These observations were supported by the regression analysis. A main effect of group at the linear term ($\beta = -2.80$, SE = 1.20, $\chi^2(1) = 5.19$, $p < 0.05$) indicates a significance difference in the overall rate of increase of looks to the target across groups. This main effect was modulated by an interaction with discourse condition at the linear term ($\beta = -1.33$, SE = 0.57, $\chi^2(1) = 5.51$, $p < 0.05$), which indicated that both L1 and L2 listeners looked more quickly to the target when it was discourse-new vs. discourse-given (L1: $\beta = -0.85$, SE = 0.08, $\chi^2(1) = 125.80$, $p < 0.001$; L2: $\beta = -1.14$, SE = 0.08, $\chi^2(1) = 187.31$, $p < 0.001$), although the effect was larger for the L2 group (steeper slopes in Figure 3.2C vs. 3.2A, more so in Figure 3.2D vs. 3.2B).

Listeners also looked more quickly to the target when it was reduced vs. unreduced, as shown by a main effect of reduction condition at the linear term ($\beta = -2.35$, SE = 0.28, $\chi^2(1) = 68.85$, $p < 0.001$; steeper slopes for circles vs. triangles). Furthermore, the peakedness of the

curve (quadratic term) was influenced by reduction condition (β = -1.34, SE = 0.28, $\chi^2(1)$ = 22.28, $p < 0.001$; sharper peaks for circles vs. triangles), although this effect interacted with discourse condition as well (β = -1.21, SE = 0.57, $\chi^2(1) = 4.55$, $p < 0.05$). This interaction reflects a sharper peak for looks to reduced vs. unreduced targets especially when the target was discourse-given (discourse-given: β = -1.94, SE = 0.36, $\chi^2(1) = 29.45$, $p < 0.001$; discourse-new: β = -0.73, SE = 0.34, $\chi^2(1) = 4.74$, $p < 0.05$; sharper peaks green vs. red lines in Figure 3.2A/3.2B than red vs. green lines in Figure 3.2C/3.2D).

Critically, each of these effects and interactions are contingent on the (critical) higher order interaction between discourse condition, reduction condition, and group, at the linear and quadratic terms (linear: β = 6.35, SE = 1.13, $\chi^2(1) = 31.44$, $p < 0.001$; quadratic: β = 4.63, SE = 1.13, $\chi^2(1) = 16.76$, $p < 0.001$). This interaction shows how groups differed in how the congruency of discourse-dependent probability and reduction influenced the overall slope (linear term) and peakedness (quadratic term) of the curve, indicated by significant three-way interactions at the linear and quadratic polynomial terms for discourse condition by reduction condition by group.

The slope of the fixation curve for both L1 and L2 listeners was influenced by discourse condition and reduction condition, although to different degrees. Follow-up regressions revealed a significant two-way interaction for both L1 (β = 3.29, SE = 0.77, $\chi^2(1) = 18.42$, $p < 0.001$) and L2 listeners (β = -3.05, SE = 0.83, $\chi^2(1) = 13.49$, $p < 0.001$). In discourse-given trials, L1 listeners did not show a significant effect of reduction condition (β = -0.41, SE = 0.48, $\chi^2(1)$ = 0.75, $p > 0.05$; no difference in slope of green vs. red lines in Figure 3.2A), whereas L2 looks to the target increased at a greater rate for reduced vs. unreduced trials (β = -4.17, SE = 0.53, $\chi^2(1)$ = 60.74, $p < 0.001$; steeper slope for green vs. red lines in Figure 3.2B). For discourse-new trials,

both L1 (β = -3.71, SE = 0.43, $\chi^2(1) = 71.19$, $p < 0.001$; Figure 3.2C) and L2 listeners (β = -1.12, SE = 0.51, $\chi^2(1) = 4.72$, $p < 0.05$; Figure 3.2D) showed significantly steeper slopes for reduced vs. unreduced trials.

The peakedness of the fixation curve differed across conditions for L2 but not L1 listeners. Follow-up regressions at the quadratic term showed that the discourse condition by reduction condition interaction was significant only for the L2 listeners (L1: β = 1.11, SE = 0.77, $\chi^2(1) = 2.10$, $p > 0.05$; L2: β = -3.52, SE = 0.83, $\chi^2(1) = 17.90$, $p < 0.001$). In discourse-given trials, the peak of the curve for L2 listeners was significantly sharper when the target was reduced vs. unreduced (β = -3.70, SE = 0.53, $\chi^2(1) = 47.95$, $p < 0.001$; sharper peak for green vs. red lines in Figure 3.2B). However, there was no effect of reduction condition for discourse-new trials (β = -0.18, SE = 0.51, $\chi^2(1) = 0.12$, $p > 0.05$; no difference in peak for green vs. red lines in Figure 3.2D).

3.2.2.3.3 ***Fixations over time: Maintaining target fixations.*** Next we turn to factors influencing how long listeners maintained looks to the target. The drop off in looks to the target was less pronounced for unreduced vs. reduced productions (more shallow decrease for triangles vs. circles) and for discourse-given vs. discourse-new trials (more shallow decrease in Figure 3.2A/3.2B vs. 3.2C/3.2D). Furthermore, L1 listeners exhibited more prolonged looks to the target (more shallow decrease in Figure 3.2A/3.2C vs. 3.2B/3.2D), especially for unreduced vs. reduced productions (more shallow decrease for triangles vs. circles in Figure 3.2A/3.2C).

Growth curve analysis. These observations were confirmed by the regression analyses. A significant main effect of reduction condition at the cubic term (β = -2.35, SE = 0.28, $\chi^2(1) = 68.85$, $p < 0.001$) reflects a more pronounced drop off in looks to the target for reduced vs. unreduced productions. Similarly, listeners maintained looks to the target longer in discourse-

given vs. discourse-new conditions, indicated by significant main effects of discourse condition at the cubic and quartic terms (cubic: $\beta = 2.69$, SE = 0.28, $\chi^2(1) = 89.59$, $p < 0.001$; quartic: $\beta = -1.33$, SE = 0.28, $\chi^2(1) = 22.25$, $p < 0.001$).

Finally, there were significant differences across groups at the quartic term. The quartic time term often reflects an asymmetry at the inflection point of curves (Mirman, Dixon, & Magnuson, 2008); in this case, it may reflect the that the L1 listeners show more prolonged looks to the target (with a flatter curve following the peak of the curve). There was a main effect of group at this term ($\beta = 0.81$, SE = 0.34, $\chi^2(1) = 5.46$, $p < 0.05$), driven by L1 listeners ($\beta = 1.11$, SE = 0.27, $\chi^2(1) = 13.53$, $p < 0.001$; L2 listeners had no significant effect: $\beta = 0.30$, SE = 0.22, $\chi^2(1) = 1.83$, $p > 0.05$). A group by reduction condition interaction at the quartic term ($\beta = 1.17$, SE = 0.57, $\chi^2(1) = 4.32$, $p < 0.05$) was again driven by L1 listeners ($\beta = 1.11$, SE = 0.38, $\chi^2(1) = 8.43$, $p < 0.01$); L2: $\beta = -0.06$, SE = 0.42, $\chi^2(1) = 0.02$, $p > 0.05$).

All other main effects and interactions failed to reach significance ($\chi^2(1) < 3.63$, ps > 0.05). A complete list of all components to the model along with regression estimates can be found in Appendix D.

Interim Summary. When considering how eye movements to the target changed over time, we found that L2 but not L1 listeners showed sensitivity to the coupling of discourse-dependent probability and reduction on the target. That is, L2 listeners showed the expected prediction effects in discourse-given conditions (e.g., faster convergence on looks to the target when it was reduced vs. unreduced), while L1 listeners did not. These results suggest differences in how L1 and L2 listeners engage predictive processing (see Table 3.4), although not in the way we would expect based on previous research in that L1 listeners did not exhibited predictive effects at all (e.g., Dahan et al., 2002). The next phase of the analysis considers eye movements

to the competitor, which complements the target results and clarifies the differences between L1 and L2 listeners.

3.2.2.4 Eye movements to the competitor in late window.

3.2.2.4.1 ***Fixations overall.*** Fixations to the competitor are shown in Figure 3.3. Looks to the competitor complemented target looks. L2 but not L1 listeners looked at the competitor more when the target was unreduced vs. unreduced (higher proportion of looks for triangles vs. circles in Figure 3.3B/3.3D). All listeners looked more to the competitor when the target was discourse-given vs. discourse-new (higher proportion of looks in Figure 3.3A/3.3C vs. 3.3B/3.3D). Finally, there were substantial differences in how listeners responded to the coupling of discourse-dependent probability and reduction on the target. Listeners looked more to the competitor when the coupling was congruent with the competitor (but not the target); the green lines (congruent coupling) were substantially higher than the red lines (incongruent coupling) across panels in the figure.

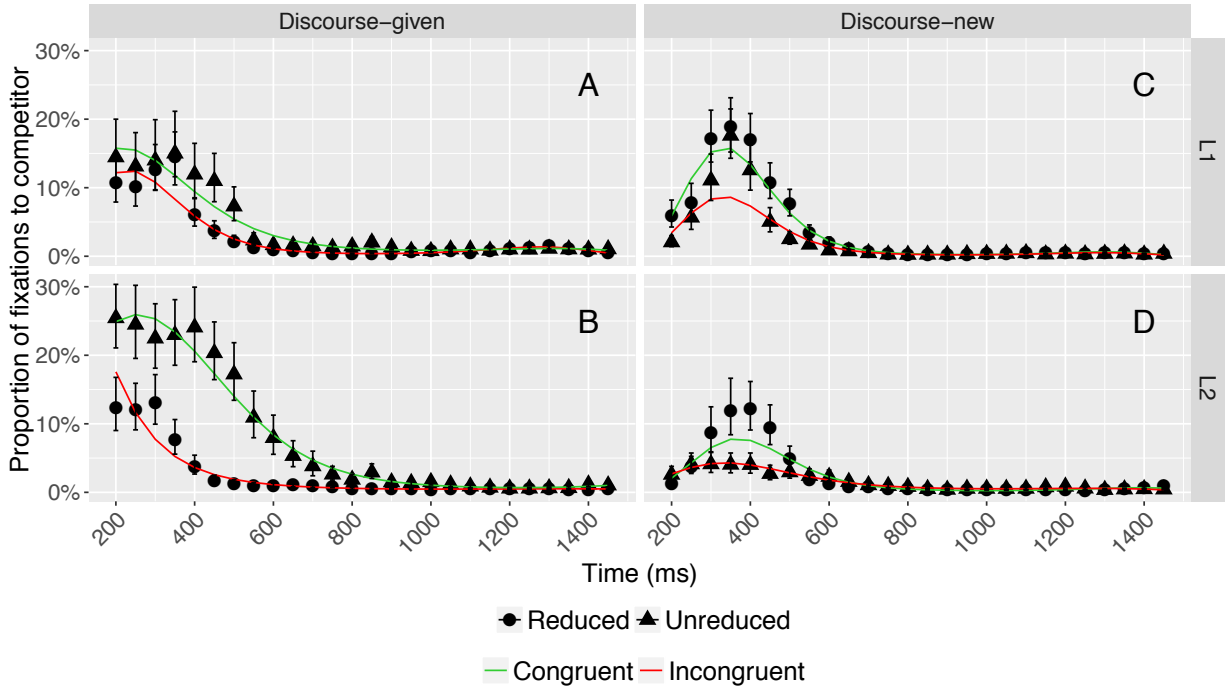

Figure 3.3. Proportion of looks to the competitor in the late window across groups (horizontal), discourse conditions (vertical), and reduction conditions (shape) in Experiment 1. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., unreduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., unreduced and discourse-new). Error bars show standard error.

Growth curve analysis. The regression results confirmed these observations. L1 and L2 listeners did not differ in the overall time spent fixated on the competitor, indicated by no significant main effect of group ($\beta = -0.08$, SE = 0.14, $\chi^2(1) = 0.37$, $p > 0.05$). Similar to the analysis of looks to the target, there were significant effects of reduction condition ($\beta = -0.35$, SE = 0.05, $\chi^2(1) = 41.69$, $p < 0.001$) and discourse condition ($\beta = 0.59$, SE = 0.05, $\chi^2(1) = 117.28$, $p < 0.001$) on looks to the competitor, where there were significantly fewer looks to the competitor in discourse-new vs. discourse-given trials (Figure 3.3C/3.3D vs. 3.3A/3.3B) and when the target was reduced vs unreduced (circles vs. triangles). A significant group by reduction condition interaction ($\beta = 0.52$, SE = 0.11, $\chi^2(1) = 22.97$, $p < 0.001$) indicates that L2 but not L1 listeners fixated more overall on the competitor when the target was reduced vs. unreduced (confirmed by

followed-up regressions: L1: β = -0.09, SE = 0.08, $\chi^2(1) = 1.39$, $p > 0.05$; L2: β = -0.61, SE = 0.08, $\chi^2(1) = 62.79$, $p < 0.001$; circles below triangles in Figure 3.3B/3.3D but not Figure 3.3A/3.3C).

Critically, the interaction between discourse condition and reduction condition was significant (β = -0.99, SE = 0.11, $\chi^2(1) = 83.77$, $p < 0.001$). Follow-up regressions indicated that listeners looked more at the competitor when the target was unreduced vs. reduced in the discourse-given condition (β = -0.85, SE = 0.07, $\chi^2(1) = 125.50$, $p < 0.001$; green lines above red lines in Figure 3.3A and 3.3B). In the discourse-new condition, looks to the competitor were higher when the target was reduced vs. unreduced (β = 0.15, SE = 0.06, $\chi^2(1) = 5.24$, $p < 0.05$; green lines above red lines in Figure 3.3C and 3.3D).

Interim Summary. The pattern of overall looks to the competitor by both groups indicated that listeners were sensitive to the coupling of discourse-dependent probability and reduction. Both groups of listeners exhibited the expected prediction effect, where they looked more to the competitor when presented with congruent coupling of discourse-dependent probability and reduction (e.g., target was discourse-given and reduced). In the analysis of overall proportions, we observed no difference in the magnitude of this prediction effect across groups. We explore differences between groups further in the analysis of the time course of looks to the competitor.

3.2.2.4.2 ***Fixations over time: Drop-off in looks to competitor.*** The overall shape of curves representing looks to the competitor complemented the target results. We begin by considering the factors that influence the drop of in looks to the competitor over time. Across groups, listeners took longer to stop looking at the competitor when the target was discourse-given vs. discourse-new (shallower drop off in Figure 3.3A/3.3B vs. 3.3C/3.3D). Listeners'

sensitivity to the coupling of discourse-dependent probability and reduction of the target was also exhibited in their looks to the competitor. When the target was discourse-given, listeners looked longer to the competitor when the target was unreduced vs. unreduced (shallower drop off for green vs. red lines in Figure 3.3A/3.3B). Similarly, when the target was discourse-new, listeners looked longer to the competitor when the target was reduced vs. unreduced (shallower drop off for green vs. red lines in Figure 3.3C/3.3D). Finally, across conditions, L1 listeners tended to prolong looks to the competitor later in time (shallower drop off in Figure 3.3A/3.3C vs. 3.3B/3.3D).

Critically, listeners' sensitivity to the coupling of discourse-dependent probability and reduction differed across groups. L2 listeners looked longer to the competitor when the coupling was congruent with the competitor (but not the target); the green lines (congruent coupling) had a substantially shallower falling slope than the red lines (incongruent coupling) in Figure 3.3B/3.3D. L1 listeners showed a similar, but smaller, effect only for trials in which the target was discourse-given, with more prolonged looks to the competitor when the target was unreduced (green lines) vs. reduced (red lines) in Figure 3.3A.

Growth curve analysis. The regression analyses confirmed these observations. The effect of the quartic component was larger for L1 vs. L2 listeners, reflected in a significant main effect of group at the quartic term ($\beta = -1.24$, SE = 0.40, $\chi^2(1) = 8.96$, $p < 0.01$). A main effect of discourse condition at the cubic and quartic terms (cubic: $\beta = -1.28$, SE = 0.28, $\chi^2(1) = 21.64$, $p < 0.001$; quartic: $\beta = 1.31$, SE = 0.28, $\chi^2(1) = 22.79$, $p < 0.001$) indicated a significant difference in how long listeners maintained looks to the competitor depending on the discourse status of the target. This main effect was modulated by an interaction with reduction condition at the cubic term ($\beta = -2.57$, SE = 0.55, $\chi^2(1) = 21.72$, $p < 0.001$). Listeners looked longer to the competitor

when the target unreduced vs. reduced in discourse-given trials (β = -1.61, SE = 0.38, $\chi^2(1)$ = 17.91, $p < 0.001$; shallower drop off for green vs. red lines in Figure 3.3A and 3.3B), and vice versa for discourse-new trials (β = 0.96, SE = 0.33, $\chi^2(1) = 8.58$, $p < 0.01$; shallower drop off for green vs. red lines in Figure 3.3C and 3.3D).

Critically, the main effect and interaction at the cubic term are contingent on the (critical) higher order interaction between discourse condition, reduction condition, and group at the cubic time term (β = 2.34, SE = 1.1, $\chi^2(1) = 4.51$, $p < 0.05$). This interaction indicates that the rate of the drop off in looks to the competitor over time differed across groups as a function of congruent pairing of discourse-dependent probability and reduction of the target.

Both L1 and L2 listeners were slower to stop looking at the competitor when probabilistic information was congruent (discourse-given, unreduced) vs. incongruent (discourse-new, reduced) with the competitor. In follow-up regressions for each group, L1 listeners showed a marginal two-way interaction of discourse condition and reduction condition at the cubic term (β = -1.40, SE = 0.78, $\chi^2(1) = 3.21$, $p < 0.08$), while the interaction was significant for L2 listeners (β = -3.74, SE = 0.78, $\chi^2(1) = 23.05$, $p < 0.001$). For L2 listeners, in discourse-given trials, listeners fixated on the competitor later in processing when the target was unreduced (β = -2.11, SE = 0.51, $\chi^2(1) = 16.87$, $p < 0.001$; shallower drop off for green vs. red lines in Figure 3.3B), whereas for discourse-new trials, this effect was observed for reduced tokens (β = 1.62, SE =0.48, $\chi^2(1) = 11.59$, $p < 0.001$; shallower drop off for green vs. red lines in Figure 3.3D). L1 listeners showed a similar (albeit smaller) effect in discourse-given trials (β = -1.10, SE = 0.56, $\chi^2(1) = 3.89$, $p < 0.05$; shallower drop off for green vs. red lines in Figure 3.3A), but no significant effect in discourse-new trials (β = 0.29, SE = 0.45, $\chi^2(1) = 0.43$, $p > 0.05$; no difference between green and red lines in steepness of drop off in Figure 3.3C).

3.2.2.4.3 ***Fixations over time: Slope and peakedness of competitor fixations.*** We now turn to factors that influence the slope and peakedness of the fixations to the competitor. Listeners began looking away from the competitor more quickly when the target was reduced vs. unreduced (sharper peak for circles vs. triangles). There was also a difference across groups in terms of how quickly listeners looked away from the competitor, where L2 listeners were somewhat delayed relative to L1 listeners (sharper peak for Figure 3.3A/3.3C vs. 3.3B/3.3D). Groups also differed in the initial rise of looks to the competitor across discourse conditions. L1 listeners looked more quickly to the competitor when the target was discourse-new vs. discourse-given (steeper slope for Figure 3.3C vs. 3.3A), while L2 listeners showed the opposite pattern (steeper slope for Figure 3.3B vs. 3.3D). Finally, the initial rise of looks to the competitor was contingent on the coupling of discourse-dependent probability and reduction. Listeners were faster to look to the competitor when the target was discourse-given and reduced vs. unreduced (steeper slope for green vs. red lines in Figure 3.3A/3.3B), and also faster when the target was discourse-new and unreduced vs. reduced (steeper slop for green vs. red lines in Figure 3.3C/3.3D).

Growth curve analysis. The regression results confirmed these observations. A significant main effect of reduction condition at the quadratic term ($\beta = 1.06$, SE = 0.28, $\chi^2(1) = 14.90$, $p < 0.001$), indicates that the competitor fixation curve was more peaked when the target was reduced vs. unreduced. The curve was also more peaked for L1 vs. L2 listeners, as indicated by a significant main effect of group at the quadratic term ($\beta = 1.05$, SE = 0.48, $\chi^2(1) = 4.49$, $p < 0.05$; sharper peak for Figure 3.3A/3.3C vs. 3.3B/3.3D). A significant interaction at the linear term showed that the groups also differed in how discourse condition influenced the slope of the curve for fixations to the competitor ($\beta = 3.22$, SE = 0.55, $\chi^2(1) = 34.15$, $p < 0.001$). Follow-up

regressions revealed that a steeper slope in discourse-new vs. discourse-given trials for L1 listeners ($\beta = 1.55$, SE = 0.39, $\chi^2(1) = 15.70$, $p < 0.001$; steeper slope for Figure 3.3C vs. 3.3A), while the opposite pattern was shown by L2 listeners ($\beta = -1.67$, SE = 0.39, $\chi^2(1) = 18.52$, $p < 0.001$; steeper slope for Figure 3.3B vs. 3.3D).

The slope of the fixation curve (linear term) differed depending on the congruence between the discourse-dependent probability and reduction of the target. This was indicated by a significant two-way interaction between discourse condition and reduction condition at the linear time term ($\beta = 2.45$, SE = 0.55, $\chi^2(1) = 19.84$, $p < 0.001$). Follow-up regressions revealed that the slope of fixations to the competitor when the target was discourse-given was steeper when the target was unreduced vs. reduced ($\beta = 1.56$, SE = 0.38, $\chi^2(1) = 16.84$, $p < 0.001$; steeper slope for green vs. red lines in Figure 3.3A/3.3B), while the slope in the discourse-new condition was steeper when the target was unreduced vs. reduced ($\beta = -0.89$, SE = 0.33, $\chi^2(1) = 7.46$, $p < 0.01$; steeper slope for green vs. red lines in Figure 3.3C/3.3D).

All other main effects and interactions failed to reach significance ($\chi^2(1) < 3.41$, ps > 0.05; see Appendix E).

Interim Summary. As looks to the competitor unfolded over time, differences in how L1 and L2 listeners engage predictive processing emerged. As the analysis of overall looks to the competitor indicated, both groups exhibited sensitivity to the coupling of discourse-dependent probability and reduction (e.g., more looks overall to the competitor when the target was discourse-given but unreduced). This indicates both groups engaged predictive processing. However, groups differed in the time course of these prediction effects, indicating subtle differences in how L1 and L2 listeners engaged predictive processing (see Table 3.4).

3.2.2.5 Interim discussion. Together, results for looks to the target and competitor in the early and late time windows point to similarities and differences in how L1 and L2 listeners use discourse-dependent probabilistic reduction as a predictive cue during speech perception.

There were widespread similarities across groups in analyses of overall looks to the target and competitor. In both windows, all listeners have an overall bias to look at new objects in the visual display, consistent with previous studies (Arnold, 2008; Dahan et al., 2002). The effect in the early window indicates that this bias begins before listeners have had sufficient time to process the phonetic input for the target noun, suggesting this bias stems from general tendencies to fixate upon novel objects rather than any sort of linguistically-driven processing. Effects of discourse condition at late time terms for both target and competitor suggest that this bias persists late in processing. A similar fixation bias was observed for looks to both the target and competitor across reduction conditions in the late window. This bias likely reflects the fact that the complete acoustic-phonetic content of the target is available earlier when reduced vs. unreduced, driving up the overall proportion of looks in the reduced condition.

Neither group exhibited prediction effects in the early analysis window. This null effect suggests that listeners do not use discourse-dependent probabilistic reduction on the determiner to make predictions about the identity of the upcoming noun.

In the late analysis window, both L1 and L2 listeners showed the expected prediction effects in overall looks to the competitor (significant discourse condition by reduction condition interaction), but the interaction for looks to the target was not significant. There were no three-way interactions with group for either set of analyses, indicating that L2 listeners did not differ from L1 listeners in whether they engaged predictive processing (or not). However, there were differences across groups in how prediction effects unfolded over time. In the condition where

both groups showed prediction effects (looks to the competitor), L2 listeners were slower to look away from the competitor when the probabilistic information cued the competitor over the target compared to when probabilistic information cued the target. This effect was indicated by a discourse condition by reduction condition interaction at the cubic term, which was modulated by an interaction with group. The higher order interaction indicated that the effect was more pronounced for L2 vs. L1 listeners, suggesting that L2 listeners maintained activation of the competitor longer in processing than L1 listeners when the probabilistic information generated strong predictions.

The results of Experiment 1 failed to replicate some critical results from previous studies with L1 listeners. Specifically, L1 listeners in previous studies looked more to the target overall in the discourse-given condition when it was reduced vs. unreduced (Arnold, 2008; Dahan et al., 2002). One major methodological divergence between the current study and existing work was that all participants in the current study had previous experience with the stimuli in our companion study. In Experiment 2, we ran a replication study with a new set of 24 L1 participants who had no previous exposure to the stimuli.

3.3 Experiment 2

3.3.1 Method

3.3.1.1 Participants. Twenty-four native listeners of English participated in this experiment (14 female; mean age: 19.08, range: 18-21). Participants were recruited from the same population as the L1 group from Experiment 1, and were all L1 speakers of English with no history of speech impairments, hearing impairments, or color blindness.

The materials, design, procedure, and additional tasks were identical to those from Experiment 1.

3.3.1.2 Data pre-processing. The same pre-processing steps as in Experiment 1 were performed to prepare the data. Twenty-six trials (4.5%) were excluded due to errors. The low recognition and post-test error rates did not merit an accuracy analysis as in Experiment 1. 39 trials (6.8%) were excluded due to trackloss. Finally, trials with *scar*/*star* as the stimulus pair were excluded (24 trials, 4.2%). Following these exclusions, 487 trials (84.5%) were included in the eye movement analysis.

Eye movement analysis. Models for analysis of eye-movements in the early and late windows included the same fixed effects with the exception of the fixed effect for group, which was eliminated. The model for the target in the late window included a fully crossed random effects structure, with by-participant random slopes for discourse condition, reduction condition, their interaction, and interactions with each polynomial term. The late window competitor model included a similar random effects structure, except with decorrelated random slopes.

3.3.2 Results

3.3.2.1 Eye movements in the early window. Fixations to the target and competitor are shown in Table 3.8. Results mirrored those observed in Experiment 1. The listeners did not exhibit prediction effects (i.e., did not show sensitivity to the coupling of discourse-dependent probability and reduction) driven by probabilistic information in the context preceding the target. Listeners showed an overall bias to look at the target when it was discourse-new and a slight bias when it was reduced vs. unreduced. Looks to the competitor were not influenced by the discourse-dependent probability or reduction of the target.

Growth curve analysis. These observations were supported by results of the regression analysis. Listeners looked more to the target in discourse-given vs. discourse-new conditions ($\beta = -0.83$, SE = 0.26, $\chi^2(1) = 8.24$, $p < 0.01$) and marginally more when the determiner was

reduced vs. unreduced (β = -0.36, SE = 0.20, $\chi^2(1) = 3.22$, $p < 0.08$). However, the interaction between discourse condition and reduction condition was not significant (β = 0.41, SE = 0.40, $\chi^2(1) = 0.84$, $p > 0.05$). Looks to the competitor were not influenced by discourse condition, the level of reduction on the determiner, or the combination between the two (all $\chi^2(1) < 1.5$, ps > 0.05).

Table 3.8. Mean proportion of looks to the target in the early time window (0-199 ms) across discourse conditions, reduction conditions, and groups in Experiment 2. Standard error in parentheses.

	Discourse-given		Discourse-new	
	Reduced	Unreduced	Reduced	Unreduced
Target	8.0% (2.4%)	11.1% (2.7%)	22.5% (3.5%)	33.9% (4.9%)
Competitor	25.0 (4.6%)	19.3% (3.5%)	12.7% (3.9%)	13.3 (3.5%)

3.3.2.2 Eye movements to the target in the late window. Fixations to the target are shown in Figure 3.4. Listeners looked more to the target overall when it was discourse-new vs. discourse-given (higher proportion of looks in Figure 3.4B vs. 3.4A). Furthermore, the time course of looks to the target differed across discourse conditions (e.g., steeper slope in Figure 3.4B vs. 3.4A). However, listeners did not look more (or more quickly) to the target when presented with a congruent vs. incongruent coupling of discourse-dependent probability and reduction (no substantial differences in overall proportion or changes over time for green vs. red lines in Figure 3.4).

Growth curve analysis. These observations were supported by the results of the regression analysis. In overall looks to the target, there was a main effect of discourse condition (β = -1.03, SE = 0.24, $\chi^2(1) = 14.05$, $p < 0.001$), where listeners looked more to discourse-new than discourse-given targets (Figure 3.4B vs. 3.4A), replicating results from Experiment 1 and

previous studies (Arnold 2008; Dahan et al., 2002). Listeners did not look significantly more to the target across reduction conditions ($\beta = 0.16$, SE = 0.17, $\chi^2(1) = 0.89$, $p > 0.05$). Finally, the reduction condition by discourse condition interaction was not significant ($\beta = -0.10$, SE = 0.34, $\chi^2(1) = 0.09$, $p > 0.05$), similar to Experiment 1 (no overall difference in green vs. red lines in Figure 3.4). While there were significant main effects of discourse condition and reduction condition at all time terms (all $\chi^2(1) > 7.03$, ps < 0.001), there were no effects consistent with discourse-dependent predictive processing in changes to the shape of the fixation curve over time (all $\chi^2(1) < 2.40$, ps > 0.05), again replicating results from Experiment 1 (no significant differences in shape of green vs. red lines in Figure 3.4).

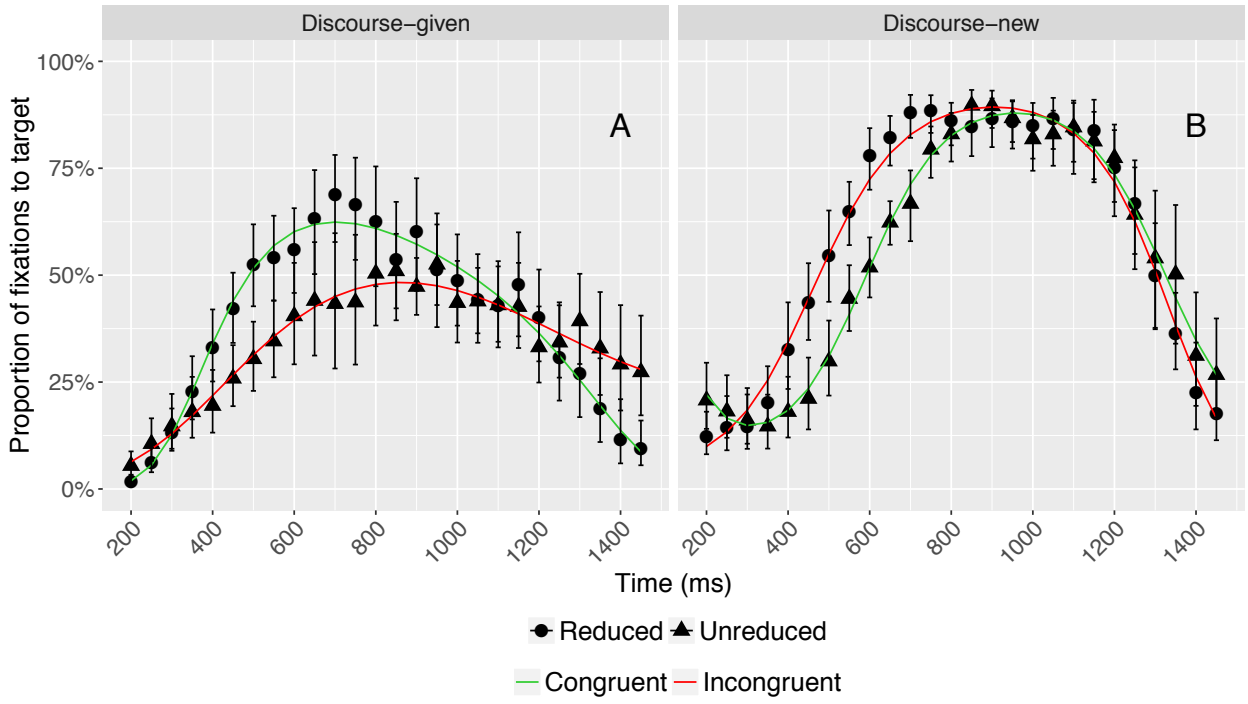

Figure 3.4. Proportion of looks to the target in the late window across discourse conditions and reduction conditions in Experiment 2. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-new). Error bars show standard error.

3.3.2.3 Eye movements to the competitor in the late window.

3.3.2.3.1 ***Fixations overall.*** Fixations to the competitor in Experiment 2 are shown in Figure 3.5. As in previous analyses, listeners looked to the competitor a higher proportion of the time when it was discourse-new vs. discourse-given (i.e., when the target was discourse-given vs. discourse-new; higher proportion in Figure 3.5A vs. 3.5B). Listeners also exhibited sensitivity to the coupling of discourse-dependent probability and reduction in terms of the overall proportion of time spent looking at the competitor. When the target was discourse-given but unreduced, listeners looked at the competitor a higher proportion of the time than when the target was reduced (green lines above red lines in Figure 3.5A).

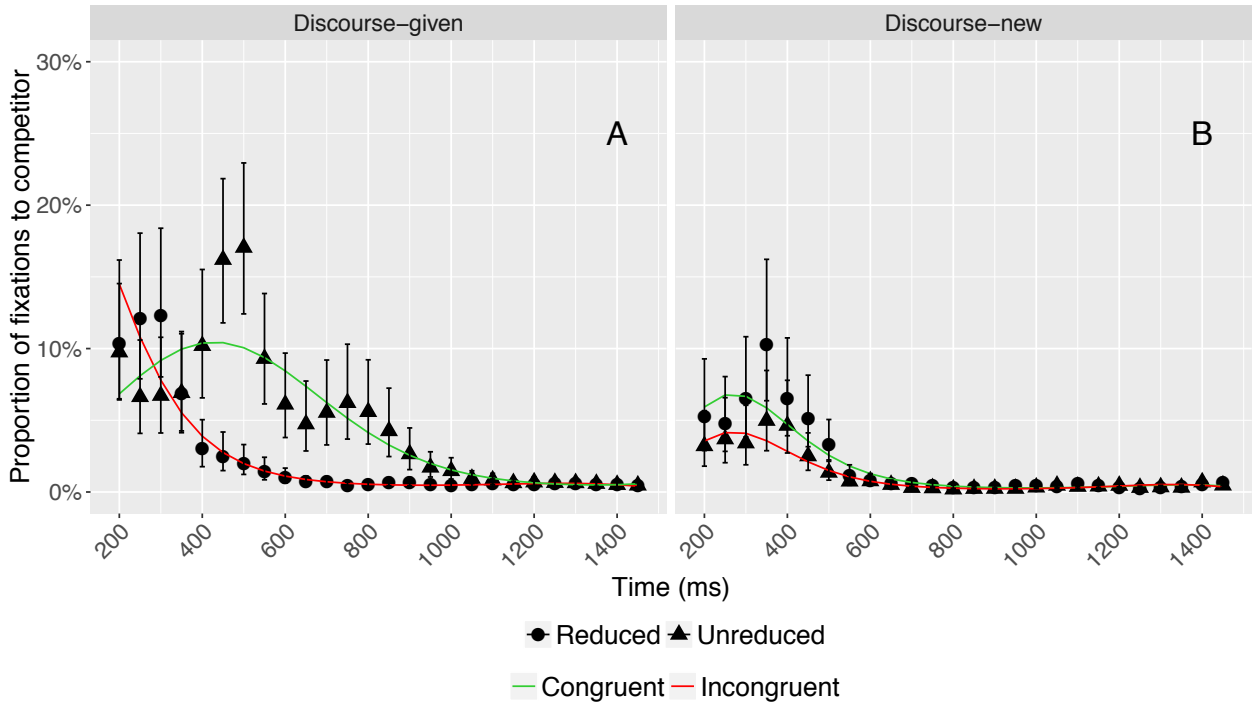

Figure 3.5. Proportion of looks to the competitor (empirical logit transformed) in the late window across discourse conditions and reduction conditions in Experiment 2. Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., unreduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., unreduced and discourse-new). Error bars show standard error.

Growth curve analysis. These observations were confirmed by the results of the regression analysis. Listeners looked more to the competitor when the target was discourse-given vs. discourse-new ($\beta = 0.83$, SE = 0.24, $\chi^2(1) = 9.92$, $p < 0.01$; Figure 3.5A vs. 3.5B), again indicating that listeners look more to the discourse-new object in the display. The main effect of reduction condition on looks to the competitor was not significant ($\beta = 0.83$, SE = 0.23, $\chi^2(1) = 2.01$, $p > 0.05$). Critically, the discourse condition by reduction condition interaction was significant ($\beta = -1.25$, SE = 0.39, $\chi^2(1) = 8.61$, $p < 0.01$), replicating the results from Experiment 1. Follow-up regressions revealed a significantly higher proportion of looks to the competitor in discourse-given conditions when the target was unreduced vs. reduced ($\beta = -0.95$, SE = 0.37, $\chi^2(1) = 5.80$, $p < 0.05$; green lines above red lines in Figure 3.5A), consistent with the results of Experiment 1. However, unlike in Experiment 1, there was no significant difference in looks to the competitor when the target was discourse-new and reduced vs. discourse-new and unreduced ($\beta = 0.30$, SE = 0.19, $\chi^2(1) = 2.34$, $p > 0.05$; no difference between lines in Figure 3.5B).

3.3.2.3.2 *Fixations over time.* Listeners also exhibited prediction effects when considering how fixations to the competitor changed over time. When the target was discourse-given and reduced, listeners maintained looks to the competitor longer in processing compared to when the target was unreduced (shallower drop off for green vs. red line in Figure 3.5A).

Growth curve analysis. This observation was confirmed by the regression analysis. As in Experiment 1, listeners were slower to stop looking to the competitor as a function of the discourse and reduction conditions, as indicated by two-way interactions at the quadratic ($\beta = 4.10$, SE = 1.25, $\chi^2(1) = 8.93$, $p < 0.01$) and cubic time terms ($\beta = -2.65$, SE = 1.16, $\chi^2(1) = 4.70$, $p < 0.05$). In the discourse-given condition, listeners continued to look to the competitor late in processing when the target was unreduced vs. reduced (quadratic: $\beta = 3.94$, SE = 1.17, $\chi^2(1) =$

9.29, $p < 0.01$; cubic: $\beta = -2.25$, SE = 0.91, $\chi^2(1) = 5.49$, $p < 0.05$; shallower drop off for green vs. red line in Figure 3.5A). During discourse-new trials, the pattern of fixations to the competitor over time were similar across reduction conditions (quadratic: $\beta = -0.16$, SE = 0.94, $\chi^2(1) = 0.03$, $p > 0.05$; cubic: $\beta = 0.40$, SE = 0.91, $\chi^2(1) = 0.19$, $p > 0.05$; no differences over time for green vs. red line in Figure 3.5B).

Listeners also showed a significant main effect of discourse-condition at the quartic term ($\beta = 1.16$, SE = 0.52, $\chi^2(1) = 4.61$, $p < 0.05$), as in Experiment 1. All other main effects and interactions failed to reach significance (all $\chi^2(1) < 2.99$, ps > 0.05).

3.3.2.4 Interim discussion. The results of Experiment 2 largely replicated those for L1 listeners in Experiment 1. There was no discourse condition by reduction condition interaction in the early time window for the target or competitor, consistent with the lack of predictive effects for the early window in Experiment 1. Furthermore, in the late window, there were no predictive effects for overall looks to the target or at any time term, again failing to replicate previous results with L1 English listeners (Arnold, 2008; Dahan et al., 2002). However, the pattern of looks to the target trends in the predicted direction. As in Experiment 1, the critical interaction was significant in the late window for overall looks to the competitor as well as at time terms indexing the drop off in looks as processing proceeds. These competitor results replicate Experiment 1 as well as other previous studies, and provide evidence that L1 speech perception is influenced by predictive processing guided by discourse-dependent probabilistic information.

As the results of Experiment 2 largely replicated the results for L1 listeners in Experiment 1, it is likely that exposure in the preceding week to the stimuli does not modulate predictive processing in this task. We therefore expect that the results for L2 listeners in Experiment 1 do not reflect previous exposure. The failure to find any effect of previous

exposure leaves open the question of why we failed to replicate previous results with L1 listeners. We expect this reflect differences in the phonetic properties of the stimuli: how the level of reduction on the determiner was elicited (emphasized vs. unemphasized text) and/or the cross-splicing of stimuli across reduction conditions. We return to this issue in the discussion below.

3.4 General Discussion

The current study investigated whether L1 and L2 listeners differ in their use discourse-dependent probabilistic reduction as a predictive cue during speech perception. We considered whether listeners generated predictions about the identity of a target noun (e.g., *candle*) based on discourse-dependent reduction concurrent with the target (i.e., reduction of the target itself) as well as reduction in the immediately preceding context (i.e., at the determiner). Our results revealed that both groups of listeners made predictions using target-concurrent information, although L1 and L2 listeners differed in how prediction effects unfolded over time. Neither group of listeners used determiner-specific information to predict the identity of the upcoming noun.

3.4.1 Prediction by L1 and L2 Listeners from Target-Concurrent Information

Predictive processing was driven by target-concurrent probabilistic information (i.e., coupling of discourse-dependent probability and reduction of the target noun) for both L1 and L2 listeners for looks to the competitor, but not the target. These predictive effects manifested both in overall looks and in how looks unfolded over time. For overall looks, there were no differences in the effects across groups. Our findings that listeners use discourse-dependent probabilistic reduction as a predictive cue replicates previous studies with L1 listeners (e.g., Dahan et al., 2002), and also show that this predictive processing is involved during L2 speech

perception. Previous studies have also observed similar overall prediction effects for L1 and L2 listeners (e.g., Dijkgraff et al., 2016), especially in situations where the listeners have shown mastery of the information in production (Hopp, 2013, 2016), as is this case for the listeners in the current study (see Study 1).

Despite no differences across groups in the overall proportion of looks, there were differences across groups in how prediction effects unfolded over time. For looks to the competitor, where both L1 and L2 listeners showed prediction effects in the analysis of overall looks, L2 listeners maintained activation of the competitor later in processing when the target was discourse-given but unreduced (incongruent coupling of discourse-dependent probability and reduction). Therefore, while L2 listeners are clearly sensitive to this coupling, and use this probabilistic information as a predictive cue during speech perception, there are subtle differences in how L1 vs. L2 listeners use predictive cues during speech perception. Our results indicate that the time course of predictive processing differs across groups, consistent with previous studies (Dijkgraff et al., 2016).

This difference in how predictive processing influences L1 and L2 speech production highlights how the efficient use of predictive information requires listeners to abandon predictions when faced with conflicting bottom-up evidence. When the bottom-up input points strongly to the target but probabilistic information is consistent with the competitor, listeners must recognize the prediction error and rely instead on the available acoustic evidence. Our results suggest that the L2 listeners found it more difficult to recover from prediction error than L1 listeners, indicated by the delayed drop off in looks to incongruent targets and congruent competitors (Kaan, 2016). In fact, this prediction deficit led L2 listeners to make marginally more recognition errors (selecting the competitor rather than the target) in the discourse-given

condition when the target was unreduced vs. reduced, consistent with these critical eye movement results.

Aside from the group difference in the time course of looks to the competitor, an additional group difference emerged in analysis of how looks to the target unfolded over time. When presented with incongruent reduction information about the discourse-given target, L2 (but not L1) listeners exhibited a more gradual increase in looks and continued to look to the target later in processing compared to trials when presented with congruent information. In this case, L2 listeners engaged predictive processing, while L1 listeners did not. Methodological choices are likely to blame for the lack of prediction effect for L1 looks to the target, as previous studies have observed these effects (Arnold, 2008; Dahan et al., 2002).

Experiment 2 considered one possible methodological divergence from previous studies (previous exposure to stimuli), but still failed to replicate previous effects (although they trended in the right direction). Cross-splicing of the stimuli across reduction conditions could also be the source of these differences across studies. We opted to cross-splice the pre-amble (*Now put*) leading up to the critical region of trials (e.g., *the candle* or *THE CANDLE*) to ensure that any response to reduction in the signal was driven by the critical region rather than the pre-amble, which also had significantly different durations across conditions. This cross-splicing (which other studies did not undertake) could have had the unintended consequence of introducing unnatural prosodic cues leading up to the critical region. Many studies have shown that the timing relations between words in an utterance have substantial impact on lexical competition and perception (e.g., Dilley & McAuley, 2008; Salverda, Dahan, Tanenhaus, Crosswhite, Masharov, & McDonough, 2007). Unnatural timing relations introduced by cross-splicing could have disrupted L1 listeners' interpretation of reduction information in the critical target region.

L2 listeners may be less sensitive to variations in timing (Baese-Berk, Morrill, & Dilley, 2016), allowing the expected prediction effects to emerge.

3.4.2 Lack of Prediction from Contextual Information

The second goal of the current study was to investigate whether listeners used discourse-dependent probabilistic reduction of the determiner as a predictive cue to the identity of the target noun. We observed no evidence that either group of listeners used the presence or absence of reduction on the determiner to make predictions about the noun. These results stand in contrast to those from previous studies, which have found that other types of discourse-dependent probabilistic information (e.g., disfluency associated with discourse-new vs. discourse-given nouns; Arnold & Tanenhaus, 2007) contribute to predictive processing during the perception of nouns (Arnold et al., 2004; Arnold et al., 2007). The lack of prediction effects based on reduction of the determiner is surprising given that production studies have found significant reduction of determiners preceding nouns with high vs. low discourse-dependent probability (Kahn & Arnold, 2012, 2015; Study 1).

There are a few possible reasons why we failed to observe prediction effects based on contextual (i.e., determiner) information. One possibility is that phonetic variation on this short function word was too subtle and ended too quickly for the information to be useful for listeners. However, this possibility is unlikely for a number of reasons. While reduced determiners were relatively short (around 48 ms on average), unreduced determiners were substantially longer (around 134 ms longer on average) and this difference should have been quite salient to listeners. Even if one could argue that this contrast was too small for listeners to track, previous eye-tracking studies have found determiner-driven effects with even subtler phonetic cues. For example, Salverda et al. (2014) found that listeners used anticipatory coarticulation within

determiner vowels to predict the identity of the upcoming target noun. Listeners were faster to look to the target when the determiner contained coarticulation from the initial consonant of the upcoming noun vs. when the determiner contained no coarticulatory cues. This indicates that the eye-tracking methodology is highly sensitive to subtle phonetic variation, and this is, therefore, not likely to be the reason we failed to find determiner-driven effects in the current study.

An alternative possibility is that the type of reduction we elicited for the current study was unnatural, and did not closely resemble the type of reduction actually produced for determiners in discourse-given vs. discourse-new contexts. This, combined with influences from cross-splicing discussed above, may have wiped out any possible predictive effects we could have observed. Future work should address methodological concerns to further test whether such information can be used as a predictive cue, and whether L2 listeners have determiner-specific deficits in the use of discourse-dependent probabilistic reduction for prediction.

3.4.3 Mechanisms Underlying Prediction

How do existing theories of speech perception, such as TRACE (McClelland & Elman, 1986) and Shortlist B (Norris & McQueen, 2008) account for prediction effects? This is a question that has undergone recent debate in the literature (for a summary, see Norris, McQueen, & Cutler, 2016), and centers around whether prediction effects arise due to interactions between different processes or interactions between different information sources.

Interactions between different processes lead to prediction when information at higher levels of representation (e.g., the lexical level) influences processing at lower levels of representation (e.g., the phoneme level) via feedback. Predictions based on lexical frequency (Dahan et al., 2001), for example, can be accounted for with prediction via feedback within the TRACE model of speech perception. The lexical representations for high frequency nouns have

high resting activation (compared to low frequency nouns). When the listener receives initial bottom-up input for a target noun, the activation of phoneme representations consistent with that input leads to activation of the corresponding lexical representations via a feed-forward flow of activation. Activation then flows backwards through the system, boosting activation of phonemes connected to recently activated lexical representations. Lexical representations with higher resting activation (due to high frequency) will send stronger feedback to the phoneme level compared to low probability representations. Selection will, therefore, be biased towards the high frequency candidate, due to higher levels of activation via this feedback loop. However, the current results are difficult to account for with this type of prediction mechanism. While discourse-dependent probability, like lexical frequency, can be encoded in terms of resting activation, in order for the level of reduction in the input to impact the feedback loop there must be additional stipulations regarding the structure of the system (see Rohde & Ettlinger, 2012, for discussion).

Alternatively, within a Bayesian framework, the results could reflect prediction resulting from interactions between information sources. From the Bayesian perspective (e.g., Kuperberg & Jaeger, 2016; Norris et al., 2016), predictions about the identity of a word constitute beliefs about which word out of a set of possible candidates is most likely (represented by probability distributions over lexical candidates, or priors). Upon encountering bottom-up input, listeners update their beliefs (via Bayes' rule) by changing prior probability distributions to accommodate new evidence. This updated set of beliefs (the posterior probability distributions over lexical candidates) reflects an interaction between different sources of information. The posterior distribution represents both previous (possibly incorrect) predictions and the current acoustic

evidence. Before lexical selection, multiple rounds of updating may occur, ultimately leading to the selection of the word with highest posterior probability.

To account for the current results, we need only to specify that reduced words with high discourse-dependent probability have higher prior probability than unreduced words with high discourse-dependent probability. High prior probabilities index strong beliefs about the identity of a word; a listener will have stronger beliefs about the identity of a word with high discourse-dependent probability if the word is reduced vs. unreduced. If prior beliefs about the identity of a word prove to be in conflict with information in the signal (i.e., there is prediction error), the updating procedure shifts these beliefs. Furthermore, within this Bayesian framework, we may be able to account for differences observed in the current study between L1 and L2 listeners.

Our results indicate that L2 listeners have strong enough beliefs (i.e., priors) about words based on discourse-dependent probabilistic information to allow them to make predictions during speech perception. However, the slow drop-off in looks to the competitor for L2 vs. L1 listeners when it (but not the target) was congruent with the probabilistic information suggests that L2 listeners' prior beliefs are stronger than their confidence in the incoming acoustic evidence (e.g., Norris & McQueen, 2008). This allows for a stronger influence of prior belief on perception behavior for L2 vs. L1 listeners, as L1 listeners have little trouble recognizing speech in their native language. This difference in the relative weighting of prior belief and confidence in bottom-up evidence across groups would lead to differences in how quickly beliefs are updated to reflect incoming acoustic evidence. This possibility could be tested using a recent technique developed by Kleinschmidt and Jaeger (2016), which estimates listeners' priors based on the statistics they are exposed to in the experiment and the statistics of their observed behavior (i.e., posteriors).

An alternative proposal (complementary to the Bayesian perspective) by Pickering and Garrod (2013) argues that prediction during comprehension is driven by a tight relationship between speech production and speech comprehension. Specifically, their proposal focuses on forward modeling. In speech production, speakers generate a production command (i.e., message that they wish to convey), which triggers a process to implement that plan (i.e., activating the semantic, syntactic, and phonological representations needed to convey the message, and executing the utterance). In parallel, a copy of the production command is made for the forward production model, which is used to generate a prediction about the utterance. By comparing a speaker's percepts of their own productions with predicted percepts of the utterance, speakers are able to monitor their own speech and adjust future production commands in cases where there are large discrepancies between actual and predicted percepts (reminiscent of Bayesian updating in the face of prediction error).

Pickering and Garrod (2013) argue that listeners are able to make predictions about others' speech because they can make predictions about their own speech. Concretely, prediction by listeners begins by covertly imitating their interlocutor's utterance and reconstructing the production command that the speaker used to produce the utterance. The listener then uses their own forward production model (based on the speaker's inferred production command) to generate a prediction about the utterance. Therefore, a listener's predictions about incoming speech are derived by their production forward model and are likely shaped, at least in part, by their own production patterns.

Some studies discussed above have reported effects consistent with this prediction-via-production idea. For example, Hopp (2013, 2016) observed that L2 listeners who demonstrated mastery of the L2 determiner system in production made L1-like predictions during perception,

while L2 listeners who used the L2 determiner system with variable accuracy did not. Study 3 investigates this relationship between the production and perception of discourse-dependent probabilistic reduction. Under Pickering and Garrod's (2013) model, we predict that speakers who produce large degrees of reduction for nouns with high vs. low discourse-dependent probability will also be able to make strong predictions based on this information in speech perception.

3.5 Conclusions

In conclusion, the current study sheds light on the role of linguistic experience in how prediction influences speech perception. These results add to existing evidence that L2 listeners can engage predictive processing, but that there are qualitative differences in the predictions made by L1 and L2 listeners. We are left with interesting questions about what might drive these differences in L1 vs. L2 predictive processing in situations where L2 listeners possess the relevant knowledge to make predictions (as shown by their production of the relevant predictive cues; Baker et al., 2011; Study 1). Future investigation of these questions ultimately refines existing theories of prediction for both L1 and L2 listener, highlighting the virtues of bilingual research as a tool for understanding language processing more generally.

CHAPTER 4

4.1 Introduction to Study 3

The relationship between speech production and speech perception has undergone much debate in linguistics research. Intuitively, there must be some relationship, otherwise it would be difficult to explain, for example, how infants come to acquire language (we must learn to produce speech sounds corresponding to the perceptual input received in our immediate environment). Despite this apparent relationship between modalities, many researchers focus specifically on either speech production or speech perception, but not both. Some studies, though, have sought to understand the nature of this relationship. These studies have investigated the strength of the coupling between production and perception, considering which representations or processes are shared across modalities.

The current study adds to existing evidence that there is some relationship between speech production and speech perception. We consider whether speech production behavior contributes to speech perception behavior and vice versa within a set of second language (L2) English speakers. These individuals participated in two experiments that investigated how probabilistic information influences processing: a speech production experiment (Study 1), followed by a speech perception experiment (Study 2) no less than one week later. We predict that individual differences between speakers in one modality will be related to individual differences in the other modality due to similarities in how probabilistic information influences processing in production and perception.

4.1.1 Perception's Influence on Production

Some evidence supporting a tight coupling between perception and production comes from shadowing tasks in which speakers hear speech and must immediately repeat what they

heard. Results from these tasks have shown that recent perception of speech (either produced by another speaker or the listener themselves) has some influence on subsequent speech production. For example, Goldinger (1998) found that words produced by shadowing a model were more perceptually similar to the model than words produced in isolation, suggesting some influence of recent perceptual experience on production behavior likely due to shared representations across modalities. In Fowler, Brown, Sabadini, and Weihing (2003), speakers produced syllables in two conditions: one in which they immediately shadowed the syllables, and another in which they always produced the same syllable (regardless of what the model produced). Fowler et al. found that speakers were not substantially slower in shadowing vs. non-shadowing conditions, which would be expected given the additional processing needed to select the correct syllable based on the model compared to the simple task of producing a pre-determined syllable. These results suggest that perceptual priming of the shadowed syllables facilitated their production. Other studies have also shown that shadowed productions reflect the fine-grained phonetic detail of the model (Nielsen, 2011; however, see Mitterer & Ernestus, 2008).

Other evidence comes from perceptual training tasks, where speaker receive some type of training on, for example, a non-native phonological, suprasegmental, or phonotactic contrast. In Bradlow, Pisoni, Akahane-Yamada, and Tokhura (1997), native Japanese participants underwent extensive perceptual training on the English /l/-/r/ contrast over the course of 45 sessions. Despite never being asked to produce this contrast during this training period, participants showed dramatic improvement in their production of this contrast (as assessed by native English listeners) following perceptual training compared to a control group that received no training. As with the shadowing results, these results indicate some sort of shared representation across modalities. However, although most participants exhibited improvements in both production and

perception, there was no correlation in the degree of improvement across modalities, suggesting individual differences between participants in the strength of the production-perception relationship.

Other studies have found similar improvements in suprasegmental production ability following perceptual training. Wang, Jongman, and Sereno (2003) observed significant improvements in English speakers' productions of Mandarin tone contrasts after perceptual training. Following training, native Mandarin listeners more successfully identified the intended tone produced by the English speakers, indicating more accurate production. Furthermore, acoustic analyses revealed that post-training productions were more similar to native Mandarin productions. In contrast to Bradlow et al. (1997), there was a significant correlation between production and perception improvements following training.

Results from a recent study (Kittredge & Dell, 2016) expose limits on the coupling of production and perception. Over a series of experiments, participants were exposed to novel phonotactic constraints (e.g., /f/ only occurs in onset position) via production of a series of tongue twisters (e.g., *kem neg feng hes*). Speakers made more speech errors violating the phonotactic constraints (illegal errors; producing *hef* instead of *hes*) relative to speech errors obeying the constraints (legal errors; producing *fes* instead of *hes*) when exposed to conflicting constraints in perception (e.g., hearing syllables with /f/ in coda position) vs. when exposed to consistent constraints across modalities. However, this transfer from perception to production only occurred for certain perception tasks. When the perception task was phoneme monitoring (listeners heard syllables following or conflicting with the production constraint and pressed a key when the target phoneme /f/ or /s/ was heard), no transfer across modalities was observed. Significant transfer was observed, though, when the perception task required inner speech (i.e.,

inwardly producing the perception syllables) or error monitoring (i.e., identifying errors in repeated sets of syllables), tasks that Kittredge and Dell argue involve speech production processes. This set of results suggests that perception may only influence production when the perception task engages production-like processes.

These studies indicate that speech perception has some impact on speech production, although there are limits to the coupling between modalities. Speakers imitate the acoustic-phonetic characteristics of recently perceived speech, and processing can be facilitated when recent perceptions overlap with current productions. However, exposure to novel perceptual information influences how speakers implement that information in production only when the production system is engaged during perception. We now turn to complementary evidence that speech production has some influence on speech perception.

4.1.2 Production's Influence on Perception

Studies of speech perception have also considered the role that production may play in perception processing. When perceptual training is uninterrupted by production, studies have observed substantial (positive) transfer from perception to production (as reviewed above; Bradlow et al., 1997; Wang et al., 2003). However, other studies have shown that the engagement of production during perceptual training blocks learning. In a recent study, Baese-Berk and Samuel (2016) trained native Spanish participants to discriminate a novel phonological contrast from Basque (/s̪/ vs. /ʃ/). One set of participants received training only in perception and showed significant improvement in their ability to discriminate /s̪/ from /ʃ/ following training. Three other groups of participants also engaged production in some way during perceptual training. One group produced the training tokens, and remained unable to discriminate the contrast following training. A group of participants who (unlike all other participants) had some

previous experience with Basque exhibited significant discrimination improvement following training with production, but showed significantly poorer discrimination than those in the perception only group. Another set of participants who produced unrelated tokens (that did not contain either phoneme in the contrast) exhibited a similar pattern of results. In all conditions when the production system was actively engaged during training (even when different representations were involved), perceptual learning suffered. However, these results suggest that the negative impact of simultaneous production training on perceptual learning can be mitigated by experience.

Engagement of the production system is not always detrimental to perception processing. In fact, some theories argue that certain perception behavior (namely, predictive processing) is driven, at least in part, by the speech production system (Dell & Chang, 2014; Pickering & Garrod, 2013). Predictive processing allows listeners to make predictions about upcoming or incoming linguistic material prior to word recognition. For example, using a visual world eye-tracking paradigm, Dahan, Swingley, Tanenhaus, and Magnuson (2000) found that native French listeners predicted the identity of nouns based on the gender marking of the previous determiner; listeners looked to pictures with feminine names (e.g., *bouteille*, bottle) upon hearing the feminine determiner but to pictures with masculine names (e.g., *bouton*, button) upon hearing the masculine determiner. If engagement of the production system is critical for this prediction to occur, then there should be some correlational relationship between prediction effects and related production ability. Recent studies of first and second language learning have documented just this relationship.

Mani and Huettig (2012) found that children (two year olds) made predictions about upcoming nouns based on semantic information earlier in the sentence (e.g., predicted *cake* upon

hearing *The boy eats the big* but not upon hearing *The boy sees the big*), replicating previous findings with adults (e.g., Altmann & Kamide, 1999). They also measured the children's comprehension and production vocabulary size (via report from parents). A series of analyses revealed that children's prediction effects were not significantly correlated with comprehension vocabulary but were correlated with production vocabulary. This indicates a strong relationship between production skill and prediction ability in children's language comprehension.

Similar effects have been observed by Hopp (2013, 2016), who tested the ability of L2 German listeners (L1 English) to predict upcoming nouns based on the gender marking on determiners (similar to Dahan et al., 2000). Prior to perception testing, Hopp (2013) assessed participants' production mastery of the German determiner system, asking whether participants consistently or variably produced correct gender marking. The results of the perception experiment revealed that L2 listeners with consistent mastery of determiner production exhibited L1-like prediction effects, while variable producers produced weaker prediction effects than L1 listeners and consistent L2 producers. These results suggest that production mastery is a pre-requisite for prediction.

Hopp (2016) investigated this possibility further with a training paradigm. Listeners completed the perception test twice, once as a pre-test and again as a post-test, with an intermediate phase to explicitly train listeners on gender assignment. In the pre-test, listeners were variable producers and did not use grammatical gender to make predictions about the upcoming noun. Following training, L2 listeners showed significant prediction effects. However, even with training, not all listeners achieved consistent mastery of the determiner system. As in Hopp (2013), listeners with variable mastery showed weaker prediction effects than listeners with consistent mastery.

4.1.3 The Current Study

In the current study, we investigate similarities in how probabilistic information influences processing across production and perception. L2 speakers of English first participated in a speech production experiment (Study 1) where they produced significant levels of discourse-dependent probabilistic reduction (i.e., reduction of discourse-given vs. discourse-new nouns). No less than one week later, the same individuals participated in a speech perception experiment (Study 2) where they exhibited significant prediction effects based on the presence of discourse-dependent probabilistic reduction. A substantial amount of individual variation in the magnitude of these effects allowed for re-analysis of the results of Study 1 including indices of prediction ability from Study 2 as independent variables, and a re-analysis of Study 2 with an index of reduction ability from Study 1.[8] The first re-analysis examines how indices of listeners' sensitivities to reduction in perception relate to reduction in production. The second re-analysis complements this by examining how indices of the influence of probabilistic information on production – the reduction of discourse-given vs. discourse-new nouns – relate to listener's sensitivity to discourse-dependent probabilistic reduction in perception. Consistent with similarities between production and perception in how probabilistic information influences processing, both analyses reveal a significant relationship between individual differences in the two modalities.

[8] While the L1 listeners exhibited considerable variation in reduction in production, there was a small range of variation in perception effects. Given the limited power of this analysis, we did not pursue it any further.

4.2 Method

Information about the participants, materials, design, and procedure for the production and perception experiments involved in this study can be found in Study 1 and Study 2. The datasets included in this analysis are identical to those used in Study 1 and Study 2.

4.2.1 Analysis

Analysis of production's influence on perception and perception's influence on production each involved two steps. In the first step, base models were built for each domain, mirroring the analyses from Studies 1 and 2. To index individual differences in production and perception ability, best linear unbiased predictors (BLUPs) for by-participant random effects of key fixed effects and interactions were extracted from these base models (more details below). These BLUPs estimate by-participant adjustments to model parameters, allowing individual participants to differ in their intercept and in the size of any fixed effects (e.g., individual differences in the magnitude of the reduction effect). Thus, BLUPs constitute a measure of individual differences in the reduction effect of Study 1 and the prediction effects of Study 2. In all results discussed below, significance tests were performed using nested model comparison.

4.2.1.1 Base production models. Linear mixed effects regressions were built with log-transformed noun duration as the dependent variable. The model included a series of control variables that all contributed significantly to model fit, including response time, lexical frequency (all continuous variables were log-transformed and centered) and a contrast-coded effect of block (block 1 vs. block 2). Additionally, models included a contrast-coded fixed effect of discourse condition (discourse-new vs. discourse-given); a significant effect of discourse condition reveals influence of discourse-dependent probability, with high probability words reduced in duration relative to low probability words. The maximal random effects structure

supported by the data was included in each model (Barr, Levy, Schreepers, & Tily, 2013), with decorrelated by-item random slopes for response time, block, and discourse condition and decorrelated by-participant random slopes for response time, block, frequency, and discourse condition. BLUPs were extracted for the by-participant random slopes for discourse condition (i.e., index of individual differences in the reduction effect).

4.2.1.2 Base perception models. Separate linear mixed effects regressions for looks to the target and competitor were built. These regressions included unweighted empirical logits of fixations aggregated within 50 ms time bins as the dependent variable. The first time bin began at 200 ms after the onset of the target nouns, as in previous studies (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). Looks to the target and competitor were both analyzed using growth curve analysis (GCA; Mirman, 2014), which uses orthogonal polynomial terms to capture non-linear changes in the fixation curve over time and allows for examinations of individual differences in eye-tracking experiments (Mirman, Dixon, & Magnuson, 2008).

Polynomial terms up to a fourth order term were included in the regressions as fixed effects and in interactions with contrast-coded fixed effects for discourse condition (discourse-new vs. discourse-given) and reduction condition (reduced vs. unreduced), which also interacted with each other. Models included random effects based on the significant prediction effects observed in Study 2. BLUPs were extracted for these random effects. For models of fixations to the target, these included by-participant random slopes for the interaction between discourse condition and reduction condition at the linear and quadratic polynomial time terms; these interactions reveal influence of discourse-dependent probabilistic reduction on how quickly listeners look to the target (linear) and then begin looking away (quadratic), with a steeper rise and sharper peak in looks to the target for discourse-given targets when reduced vs. unreduced.

For models of fixations to the competitor, this included by-participant random slopes for the discourse condition by reduction condition interaction at the cubic time term and on the intercept (i.e., the overall effect); these interactions reveal influence of discourse-dependent probabilistic reduction on the overall proportion of looks to the competitor and how long listeners looked to the competitor (cubic), with a higher proportion and slower drop-off in looks when the target was discourse-given and reduced vs. discourse-given and unreduced.

4.2.1.3 Models of perception's influence on production. To analyze how sensitivity to discourse-dependent probabilistic reduction in perception influenced the ability to reduce in production, BLUPs for significant prediction effects were added to the base production model as continuous (centered) fixed effects, in interactions with discourse condition, and were included as by-item random slopes in the random effects structure. BLUPs were candidates for this analysis if there were involved in significant prediction effects in the analysis of Study 2 (see previous section). In addition, because the interaction at the linear term for looks to the target was not significant in the base perception model for this analysis, the BLUPs for the linear term were not included in the production model.

The analysis below focuses on critical effects for evaluating the hypothesis that perception influences production, which were interactions of BLUPs (i.e., measures of listeners' sensitivity to discourse-dependent probabilistic reduction as a predictive cue) from the perception model with reduction effects observed in Study 1 (all other model output can be found in Appendix H).[9] If there are similarities in the processing of probabilistic information

[9] The two sets of BLUPs from the perception competitor model were moderately correlated (ρ = 0.452). The results below held when excluding the BLUPs for the overall interaction, leaving only the BLUPs from the target model and from the cubic interaction in the competitor model.

across production and perception, then participants who exhibited strong prediction effects in perception should also produce larger reduction effects than those participants that exhibited weak effects in perception. This prediction will manifest as significant interactions between any set of BLUPs and discourse condition.

4.2.1.4 Models of production's influence on perception. To analyze the relationship between a speakers' reduction ability in production and their use of discourse-dependent reduction as a predictive cue in perception, BLUPs were extracted from base production models for the discourse condition effect. These BLUPs were included as a centered, continuous fixed effect, and entered in interactions with discourse condition, reduction condition, and the polynomial time terms.

The analysis below focuses on critical effects for evaluating the hypothesis that production influences perception, which includes interactions of BLUPs from the production model with prediction effects (all other model output can be found in Appendices I and J). Prediction effects constitute interactions between discourse condition and reduction condition overall and at polynomial terms, reflecting facilitation of perception by the appropriate coupling of discourse-dependent probability and reduction. If there are similarities in the processing of probabilistic information across modalities, then good reducers (who produce the largest degrees of discourse-dependent probabilistic reduction) should exhibit strong prediction effects while bad reducers should show weaker prediction effects. This prediction will manifest as a significant interaction between BLUPs and prediction effects (interactions involving discourse condition and reduction condition).

4.3 Results

4.3.1 Perception's influence on production

Overall word durations were influenced by BLUPs for the prediction effect at the cubic term from the competitor model (β = -0.02, SE = 0.01, $\chi^2(1) = 5.83$, $p < 0.05$), indicating that listeners who exhibited the most pronounced prediction effects on the cubic term for looks to the competitor also tended to have shorter durations overall in production. The other two perception BLUPs did not have a significant effect on overall word durations ($\chi^2(1) < 1.53$, ps > 0.05). Critically, there was a significant interaction between discourse condition and the cubic competitor BLUPs (β = 0.007, SE = 0.003, $\chi^2(1) = 5.66$, $p < 0.05$). Neither of the other two perception BLUPs significantly influenced the magnitude of the condition effect ($\chi^2(1) < 1.80$, ps > 0.05).

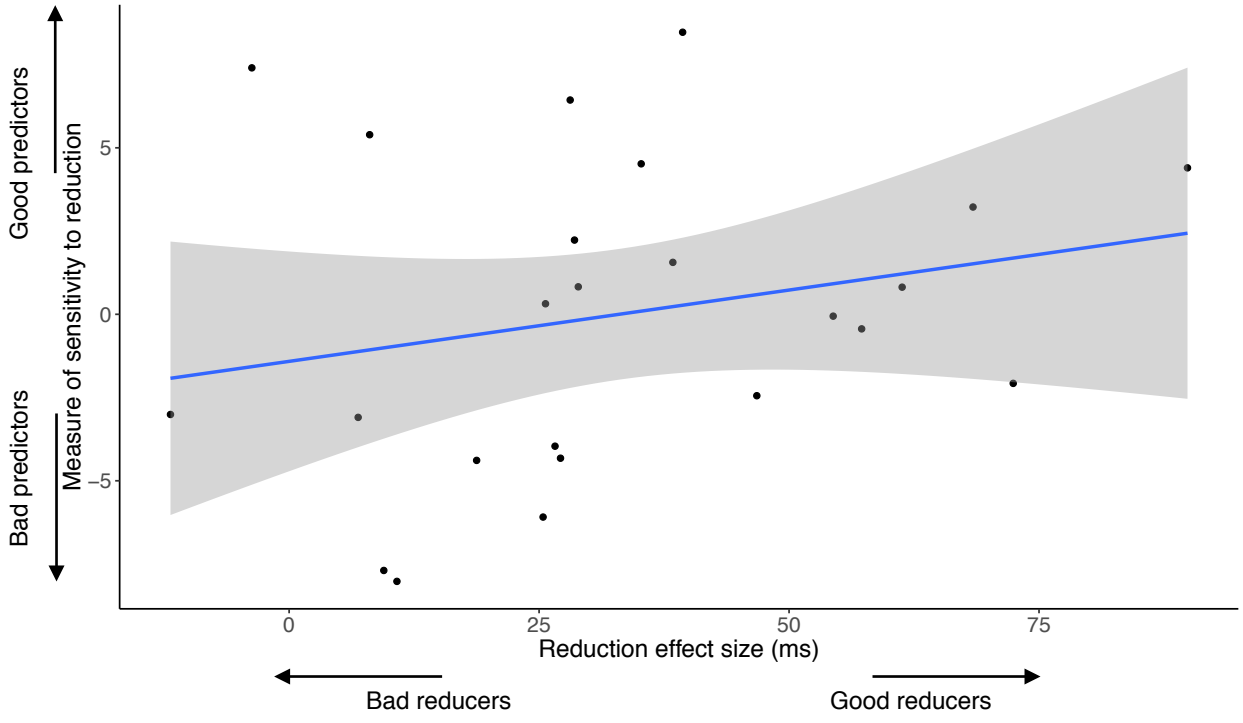

Figure 4.1. Reduction effect size (discourse-new – discourse-given word durations) by BLUPs from reduction effect at cubic term for looks to the competitor in the perception experiment (an

index of listeners' sensitivity to reduction as a predictive cue). Regression line shows simple linear regression with 95% confidence interval.

This interaction can be seen in Figure 4.1, which plots the reduction effect size (durations of discourse-given productions subtracted from durations of discourse-new productions) against the measure of sensitivity to reduction that was significant in the above regression analysis (i.e., the cubic competitor BLUPs). Listeners with high values for this sensitivity measure had larger prediction effects in perception (i.e., "good predictors") compared to those with smaller values who had smaller prediction effects (i.e., "bad predictors"). This figure, and the significant interaction, indicates a linear relationship between the prediction measure and the reduction effect, where good predictors also produce large reduction effects but bad predictors produce small reduction effects. For example, the two participants who have the lowest levels of sensitivity to reduction in perception (the worst predictors), also produce very little reduction in production (among the worst reducers). It should be noted that although this relationship was statistically significant, it is relatively weak; this indicates that other individual differences between speakers (above and beyond than their prediction behavior) likely influence the degree to which they reduce nouns with high vs. low discourse-dependent probability.

4.3.2 Production's influence on perception

4.3.2.1 Looks to the target. To examine the influence of production behavior on listeners' prediction ability, we focus on interactions between production BLUPs (the index of reduction ability from the production model) and the prediction effect at the linear and quadratic terms. The influence of the prediction effect on the slope of the fixation curve (linear term) was significantly influenced by production behavior, as shown by a three-way interaction between BLUPs, discourse condition, and reduction condition at the linear term ($\beta = 0.19$, SE = 0.04, $\chi^2(1) = 26.57$, $p < 0.001$). The interaction at the quadratic term was not significant ($\beta = -0.0004$,

SE = 0.04, $\chi^2(1) = 0.0001$, $p > 0.05$). In order to investigate this interaction, participants were assigned to two groups based BLUPs values; "good reducers" had a BLUP value greater than the mean (0), and "bad reducers" had values below the mean. Figure 4.2 shows fixations to the target across conditions and groups.

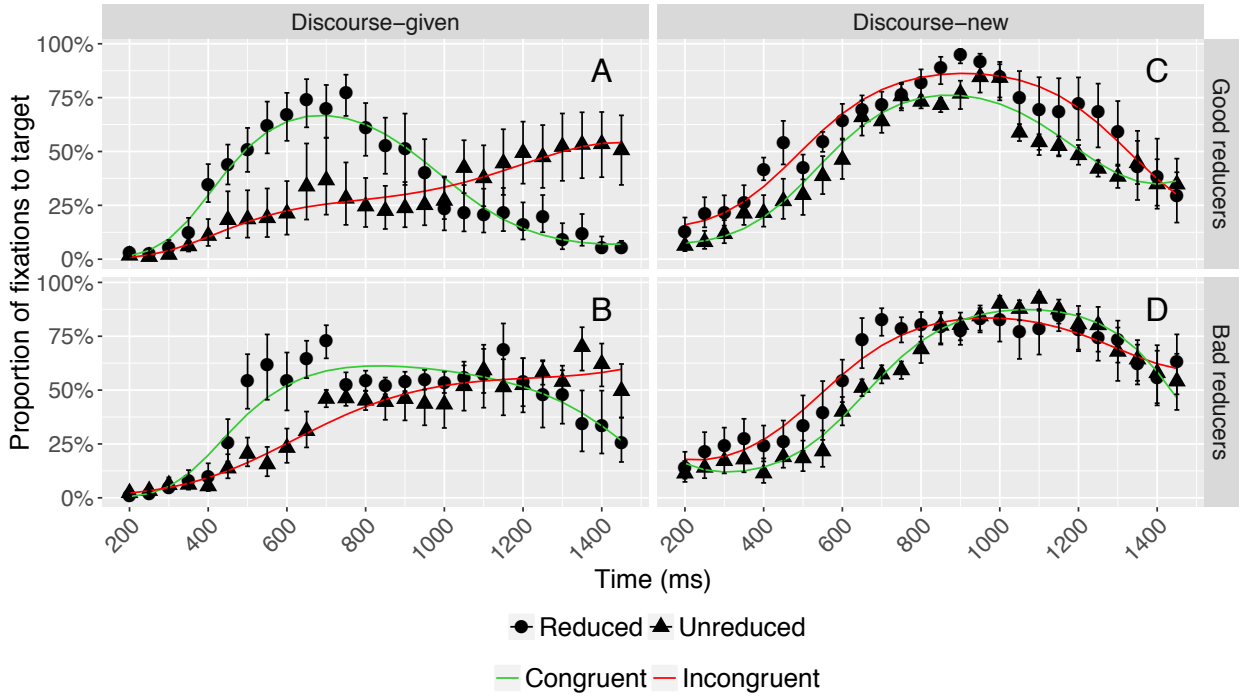

Figure 4.2. Proportion of looks to the target separated by group (horizontal), discourse conditions (vertical), and reduction conditions (shape). Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and discourse-given) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-new). Error bars show standard error.

Prediction effects manifest on the linear term of the growth curve model when listeners increase looks to the target significantly faster when presented with congruent probabilistic information (e.g., a reduced, discourse-given target) vs. incongruent probabilistic information (e.g., an unreduced, discourse-given target). The interaction at the linear term indicates that listeners' ability to use discourse-dependent reduction to make predictions during speech

perception is related to their ability to produce the same sort of reduction. Follow-up regressions were run for each of the groups described above.

For good reducers, listeners increased looks to the target more quickly when it was discourse-given and reduced (green line) vs. discourse-given and unreduced (red line; Figure 4.2A). In this significant prediction effect for good reducers, listeners quickly identified the target when presented with congruent probabilistic information, but exhibited uncertainty about the identity of the target when presented with incongruent information. In contrast, bad reducers did not show significant prediction effects discourse-given trials (Figure 4.2B). Consistent with Study 2, neither group shows differences in the slope of the fixation curve across reduction conditions when the target was discourse-new (Figures 4.2C/D).

The pattern of results shown in Figure 4.2 reveals that prediction effects were observed for listeners with large reduction effects in production but not for listeners with small reduction effects. This observation was supported by follow-up regression analyses for each group of listeners. For listeners in the good reducer group, there was a significant two-way interaction between discourse condition and reduction condition at the linear term ($\beta = 0.29$, SE = 0.05, $\chi^2(1) = 29.60$, $p < 0.001$); the interaction was not significant for bad reducers ($\beta = 0.006$, SE = 0.05, $\chi^2(1) = 0.01$, $p > 0.05$). The significant interaction for good reducers reflected a significant main effect of reduction condition at the linear term in discourse-given trials ($\beta = -0.65$, SE = 0.07, $\chi^2(1) = 80.09$, $p < 0.001$) but not discourse-new trials ($\beta = -0.06$, SE = 0.07, $\chi^2(1) = 0.78$, $p > 0.05$). These results indicate that when probabilistic information had a strong impact on production processing (for good reducers) it also had a substantial influence on perception processing, leading to strong prediction effects.

4.3.2.2 Looks to the competitor. For competitor looks, prediction effects manifest when listeners are slower to look away from the competitor when presented with probabilistic information that is congruent with the competitor (e.g., an unreduced, discourse-given target) compared to incongruent probabilistic information (e.g., a reduced, discourse-given target). This is reflected in changes to the cubic term in the growth curve model across discourse and reduction conditions. These prediction effects were influenced by participants' production behavior; there was a significant interaction between production BLUPs and the prediction effect at the cubic term ($\beta = 0.13$, SE = 0.04, $\chi^2(1) = 13.11$, $p < 0.001$). Figure 4.3 shows looks to the competitor across conditions and groups. To investigate this interaction, separate regressions were run for a group of "good reducers" (listeners with BLUP values above the mean) and a group of "bad reducers" (listeners with BLUP values below the mean).

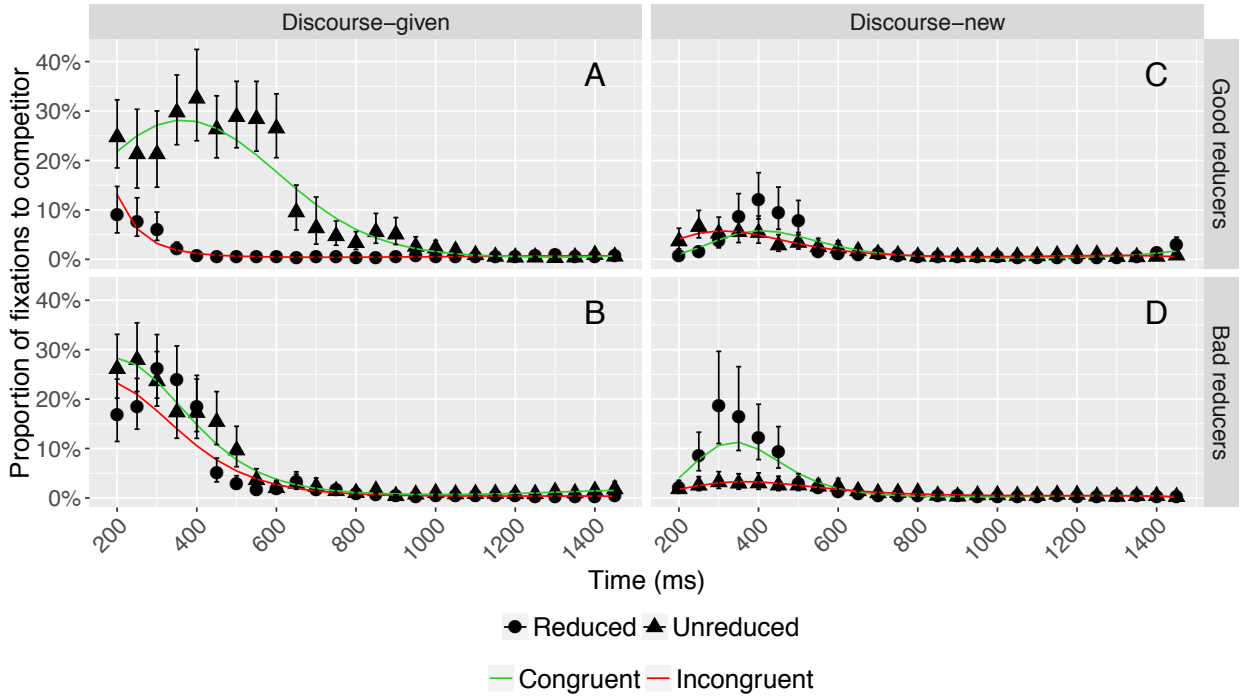

Figure 4.3. Proportion of looks to the competitor separated by group (horizontal), discourse conditions (vertical), and reduction conditions (shape). Lines show growth curve model fit, with green lines corresponding to congruent reduction and discourse conditions (e.g., reduced and

discourse-new) and red lines corresponding to incongruent conditions (e.g., reduced and discourse-given). Error bars show standard error.

Good reducers showed strong effects of prediction; in discourse-given trials (Figure 4.3A) they were slower to look away from the competitor when the target contained congruent (green line) vs. incongruent probabilistic information (red line). In contrast, bad reducers showed little prediction effect when the target was discourse-given (Figure 4.3B). Neither good nor bad reducers appeared to look substantially longer at the competitor when the target was discourse-new and reduced vs. unreduced (Figure 4.3C/D).

The pattern of results shown in Figure 4.3 reveals that only good reducers exhibited prediction effects. These observations were supported by regression analysis, which revealed that listeners in the good reducer group showed a two-way interaction of discourse condition by reduction condition at the cubic term ($\beta = 0.33$, SE = 0.05, $\chi^2(1) = 37.73$, $p < 0.001$), but the bad reducer group did not ($\beta = 0.04$, SE = 0.05, $\chi^2(1) = 0.57$, $p > 0.05$). In discourse-given trials, good reducers continued to look to the competitor later in processing when the target was unreduced vs. reduced, as shown by a main effect of reduction at the cubic term ($\beta = -0.39$, SE = 0.06, $\chi^2(1) = 35.55$, $p < 0.001$; more gradual drop off for green vs. red line in Figure 4.3A). In contrast, a main effect of reduction at the cubic term in the discourse-new condition ($\beta = 0.27$, SE = 0.07, $\chi^2(1) = 14.33$, $p < 0.001$) indicates that good reducers showed a slower drop off in looks to the competitor when the target was discourse-new and reduced (green line in Figure 4.3C) vs. discourse-new and unreduced (red line in Figure 4.3C).

4.4 General Discussion

The current study investigated the relationship between speech production and perception behavior by L2 speakers of English. The results of Study 1 and Study 2 revealed individual differences in how probabilistic information influences processing during speech production and

perception. We considered whether these individual differences are related across modalities. In the first analysis, we asked whether individual differences in prediction effects from the speech perception task of Study 2 related to listeners' ability to reduce word durations during the speech production task of Study 1. The results of this analysis revealed that listeners who make good predictions in speech perception also produce large degrees of reduction in production. In the second analysis, we asked whether individual differences in reduction effects from the speech production task of Study 1 related to listeners' ability to predict during speech perception. The results of this analysis indicated that good reducers exhibited strong prediction effects during speech perception, while bad reducers exhibited weaker or no prediction effects. Together, these results indicate a strong link between production and perception in this particular domain (i.e., how probabilistic information influences processing).

These results are in line with a series of other studies that have investigated the relationship between production and perception. As in other prediction studies (Hopp, 2013, 2016), we observed that the ability to make predictions during speech perception depends (at least in part) on one's ability to produce the cues on which predictions are based. This result is predicted by theories arguing that prediction involves some sort of production processing (Dell & Chang, 2014; Pickering & Garrod, 2013). Furthermore, we observed that the ability to produce discourse-dependent probabilistic reduction relates to one's ability to make predictions using this reduction. This result is broadly consistent with Kittredge and Dell (2016), who demonstrated that relationships between production and perception manifest when perception involves engagement of the production system (which we assume is the case in studies of prediction). Our findings extend previous work by showing a mutual influence between production and perception within the same set of individuals and by investigating this relationship in terms of

the processing of probabilistic information. We found evidence of a tight coupling between production and perception in this particular domain, while other studies have found a flexible coupling (Bradlow et al., 1998) or even a parasitic relationship between the two (Baese-Berk & Samuel, 2016). Future research should continue to investigate this relationship by carefully considering what processes or representations may be shared across modalities, so that we can gain a more complete understanding of this relationship and under what conditions we should see transfer across modalities.

As a general framework for understanding these results, we explore existing theories of prediction (Dell & Chang, 2014; Kuperberg & Jaeger, 2016; Norris, McQueen, & Cutler, 2016; Pickering & Garrod, 2013), which are useful for understanding how probabilistic information can influence processing in both production and perception. While the details of these accounts differ in nuance, they share the common thread that prediction stems from leveraging existing knowledge (e.g., forward production models (Dell & Chang, 2014; Pickering & Garrod, 2013); prior beliefs (Kuperberg & Jaeger, 2016; Norris et al., 2016)) to make sense of new input. These theories can easily account for our findings that listeners who produce large levels of reduction in production also make strong predictions during perception. For example, under Pickering and Garrod's model, during perception an individual's forward model generates predictions about how they would produce the incoming input. Therefore, hearing a reduced production for a noun with high discourse-dependent probability matches the predictions that a good reducer would generate in that context. On the other hand, a bad reducer reduces discourse-given vs. discourse-new words less consistently, so they will make less consistent predictions in the same context. Similarly, Bayesian models (e.g., Norris et al., 2016) argue that listeners have a strong prior belief that words with high discourse-dependent probability will be reduced. Bottom-up input

that conforms to that belief allows listeners to make strong predictions about the identity of the incoming word. Good reducers will have stronger beliefs compared to bad reducers and will, therefore, make stronger predictions overall than bad reducers.

Our findings that prediction ability influences production of the cues that drive prediction are also compatible with these accounts. We propose that prediction ability influences production via perceptual learning over the course of an individual's lifetime (see also Dell & Chang, 2014). When listeners make predictions (using their forward production model or prior beliefs), the outcome of these predictions is compared against the actual input (in a monitoring process or via Bayesian updating). If the predictions are wrong (i.e., do not match the actual input), this prediction error is used to update the forward production model or the prior beliefs used to generate the predictions. In the Pickering and Garrod (2013) perspective, changes to the forward production model have consequences for future productions. When speakers are later in situations where they should reduce a discourse-given word, the prediction generated by the forward production model (which now reflects previous perception prediction error) may differ from the actual utterance and lead to another prediction error. This production-based prediction error allows the speaker to adjust future production commands to account for this discrepancy. In the Bayesian perspective (e.g., Norris et al., 2016), the posterior beliefs derived after error-based updating constitute the prior beliefs used during subsequent processing and, thus, influence any subsequent productions. Under this proposal, speakers learn to produce discourse-dependent probabilistic reduction by a similar process as speakers learn to produce novel phonological contrasts (e.g., Bradlow et al., 1997).

Future work could investigate this hypothesis in a perceptual training paradigm. This hypothesis predicts that with enough perceptual training (for example, completing our perception

task a number of times), bad reducers could learn to reduce discourse-given vs. discourse-new words to a larger degree. This sort of training was not possible in the current study, as all participants completed the production task followed by the perception task. This hypothesis also predicts that perceptual training reinforcing the opposite relationship between discourse-dependent probability and reduction (i.e., high probability words tend to be unreduced, and low probability words reduced) should create bad reducers from good reducers. How quickly this sort of shift could possibly occur is an empirical question. However, other studies have shown that listeners quickly adapt to the statistics within an experiment (e.g., Fine, Jaeger, Farmer, & Qian, 2013) and adjust predictive processing over the course of the experiment based on the new statistics (e.g., Hopp, 2016). These questions should be considered by future studies.

4.5 Conclusions

In conclusion, the current study provides evidence of a relationship between speech production and perception in terms of how probabilistic information influences processing. Specifically, these results speak to how speech production influences prediction during speech perception and how prediction influences production of the cues used to make predictions. By considering individual differences in production and perception behavior in parallel, we can elaborate on existing theories (e.g., Norris et al., 2016; Pickering & Garrod, 2013) and gain a fuller understanding of how speakers learn to reduce high vs. low probability targets and how listeners learn to predict based on this reduction.

CHAPTER 5

The goal of this dissertation was to better understand how probabilistic information influences processing during speech production, speech perception, and across modalities. To achieve this goal, we examined the speech behavior of individuals with different levels of linguistic experience. The final chapter of this dissertation summarizes the major findings of each of the three studies within. We consider the implications of these findings for understanding the role of phonetic variation during production, perception, and communication more generally, and we discuss possibilities for future work.

5.1 The Shape of Phonetic Variation Depends on Linguistic Experience and Word Class

In Study 1, we observed that discourse-dependent probabilistic information has a comparable impact on L1 and L2 content word processing; both groups of speakers reduced the duration of content words with high vs. low discourse-dependent probability. Similarly, there was no difference across groups in how response times (an index of planning speed) were influenced by probabilistic information. However, this probabilistic information had distinct influences on function word processing across groups; L1 but not L2 listeners produced reduced durations for function words preceding nouns with high vs. low discourse-dependent probability. This difference between groups in the production of function words indicates that L2 speakers have a particular deficit in the processing of probabilistic information for function words. By virtue of this deficit, we have support for the hypothesis that the reduction produced by L1 listeners likely stems from probability inheritance rather than priming-based facilitation; function words inherit the probability of the content words they precede as long as function word planning proceeds in parallel with content word planning.

These results have implications for our understanding of speech production outside the limited scope of single word production, which has been the typical focus of work in psycholinguistics (e.g., Levelt, Roelofs, & Meyer, 1999). Other studies using the same event description paradigm from Study 1 similarly expand our understanding of the dynamics of the speech production system under carefully controlled but more ecologically-valid circumstances than single word picture naming or paragraph reading. As in previous studies (e.g., Jurafsky, Bell, Gregory, & Raymond, 2001), our results indicate that function word processing is highly sensitive to the probabilistic information associated with surrounding words, indicating that function word production is best understood in the context of multi-word productions.

Future work should consider how different types of structural relationships between function and content words influences phonetic variation. Our findings with determiner-noun productions could be attributed to the modifier relationship between determiners and noun phrases. It could be the case that probability inheritance is restricted to this type of structural relationship, and we would not expect to see reduction of function words that are not modifiers to high probability content words (e.g., verbs providing probability to pronouns in specifier position). Alternatively, c-command relations could be the critical structural characteristic for understanding these effects. Both determiners and pronouns c-command the content words that follow them. If probability inheritance depends on this relationship, we should expect to see reduction in both cases. Future studies should expand the current work by considering other structural relationships between different lexical categories (e.g., between pronouns and verbs, between adjectives and nouns, etc.).

Another important avenue for further inquiry concerns how differences in syntactic or morphological structure across a bilingual's L1 and L2 may influence the types of effects we

observed in Study 1. Our L2 speakers were L1 speakers of Mandarin, which does not have determiners. Future studies should attempt to replicate our findings with, for example, L1 German speakers, who have experience producing determiners in a language with similar prosodic structure as English. The difficulties experienced by our Mandarin speakers may be mitigated by the experience German speakers have producing determiners in their L1. Additionally, German speakers may have more experience with overlap in the scope of determiner and noun planning (argued in Study 1 to be necessary for probability inheritance to occur). Because German determiners are marked for gender, determiner planning in German necessarily overlaps in time with noun planning so that the determiner matching in gender with the noun can be selected (e.g., Miozzo & Caramazza, 1999). Comparing L1 German speakers (in English) and L1 speakers of a language with non-gender-marked determiners, could dissociate whether mere experience with determiners is enough to alleviate L2-specific difficulties. Such future work will allow us to better understand how syntactic structures influence the phonetic properties of productions.

Future research should investigate communicative interactions between L1 and L2 interlocutors in order to gain a better understanding of how speakers take certain characteristics of the listener into account (i.e., listener modeling) when producing discourse-dependent phonetic variation. That is, do L1 speakers produce qualitatively different variation (e.g., more or less reduction) depending on the language background of their interlocutor? A series of recent studies (e.g., Buz, Tanenhaus, & Jaeger, 2016; Hazan & Baker, 2011; Rosa, Finch, Bergeson, & Arnold, 2015) have found that speakers track listener responses over the course of a conversational exchange, for example, noting when listeners make perception errors. Following these errors, speakers adjust their productions in order to avoid future listener error. For example,

Hazan and Baker (2011) found that speakers were able to infer listener deficits, such as the inability to perceive pitch when hearing vocoded speech, and make specific modifications to their productions to increase the intelligibility of their speech for that particular listener (e.g., increase speech rate but not pitch range).

Returning to the production of discourse-dependent probabilistic reduction, if L1 interlocutors begin a conversation with the belief that an L2 interlocutor will have difficulties with the language, they may produce less reduction of nouns with high vs. low discourse-dependent probability in order to maximize redundancy in the signal for the L2 listener. Over the course of the conversation, if L1 interlocutors observe that the L2 speaker actually produces L1-like levels of discourse-dependent probabilistic reduction, they may modify their speech accordingly and produce larger levels of reduction compared to the beginning of the conversation. However, if probabilistic information influences L1 and L2 processing in different ways (e.g., for function words), leading to non-L1-like reduction from the L2 speaker, L1 speakers may not begin reducing function words as they would when speaking to an L1 interlocutor. A study such as this would shed light on how rapidly linguistic experience (in this case, experience with the variation produced by L2 speakers) influences how probabilistic information impacts processing over the course of a conversation. Furthermore, future work should attempt to understand the degree to which L2 speakers can tailor their productions to the needs of their interlocutors.

5.2 Phonetic Variation Benefits the Listener, Regardless of Linguistic Experience

In Study 2, we found that both L1 and L2 listeners engaged predictive processing during speech perception. Both groups of listeners made predictions such that target recognition was facilitated when the probabilistic information was congruent vs. incongruent with the incoming

input. However, L1 and L2 listeners differed in how these predictions unfolded over time. Specifically, L2 listeners made strong predictions about the identity of the target but took longer than L1 listeners to abandon these predictions when they conflicted with the bottom-up input. Neither group of listeners made predictions about the target upon hearing discourse-dependent probabilistic reduction in the context immediately preceding the target. These results indicate that L2 listeners are capable of engaging predictive processing, but make qualitatively different predictions than L1 listeners likely due to differences in linguistic experience across groups.

The results of Study 2 have implications for characterizing how listeners make predictions during speech perception. While perceiving speech, listeners leverage previous experience (e.g., prior beliefs from the Bayesian perspective; Kuperberg & Jaeger, 2016; Norris, McQueen, & Cutler, 2016) in order to make sense of new input. By applying previous experience (e.g., that words with high discourse-dependent probability tend to be reduced in duration), listeners can make predictions about the identity of upcoming words in the face of ambiguous input. However, because L1 and L2 listeners necessarily possess different levels of experience with a language, they also apply this experience in distinct ways when making predictions during speech perception. The differences observed in Study 2 in how predictions made by L1 and L2 listeners unfolded over time suggest subtly distinct experiences across groups in phonetic variation conditioned by discourse-dependent probability in L1 speakers of English.

Future research should consider whether L1 and L2 listeners engage predictive processing differently when presented with phonetic variation produced by L2 speakers of English. In a previous study investigating the use of disfluency associated with high vs. low probability (lexical frequency) referents, L1 listeners predicted the identity of the target noun

when disfluencies were produced by L1 but not L2 speakers (Bosker, Quené, Sanders, & de Jong, 2014). Bosker and colleagues argued that L1 listeners did not make predictions from L2 speech because they attributed the disfluencies to L2-specific speech production difficulties, rather than to a systematic relationship between probability and production difficulty (i.e., more disfluencies associated with low frequency referents; see also Arnold & Tanenhaus, 2007, for a similar relationship between discourse-new referents and disfluency). When L1 listeners deduce that the information presented to them is unreliable, they are unlikely to use that information as a cue for prediction (e.g., Hopp, 2016).

However, in contexts where L2 speakers are producing fluent, meaningful phonetic variation of the type used in Study 2, L1 listeners may be more likely to trust this variation and use it to make predictions. L1 listeners' ability to do so likely depends on their past experience with L2 speakers of English. If L1 listeners approach speech perception with weak prior beliefs that high probability words are reduced in L2 speech, they are unlikely to use this phonetic variation as a predictive cue. However, given the results of Study 1, L1 listeners may have encountered L2 speakers that reliably reduce words with high vs. low discourse-dependent probability. If this is the case, then L1 listeners likely have strong priors for L2 speech similar to those for L1 speech, allowing them to engage predictive processing. A study such as this would complement Study 2 of this dissertation nicely by considering the role of another type of linguistic experience – experience with L2 speech – during speech perception.

By investigating how L1 listeners make predictions based on L2 speech and vice versa, we can better understand the role that variation plays during communication. Study 2 takes the first step towards characterizing how linguistic experience comes into play during

communication, and how differences in linguistic experience may impede listeners' ability to make communication as efficient as possible by engaging predictive processing.

5.3 Production and Perception of Variation is Related within Individuals

In Study 3, we found that individual differences in an L2 speaker's ability to reduce nouns with high vs. low discourse-dependent probability was related to their ability as a listener to make predictions based on this probabilistic information. The complementary relationship was also observed; listeners who made strong predictions based on discourse-dependent probabilistic information also produced large degrees of reduction for nouns with high vs. low discourse-dependent probability. These results have implications for our understanding of the relationship between speech production and speech perception in terms of how probabilistic information is processed across modalities. We propose that engaging prediction constitutes perceptual learning; robust engagement of prediction during speech perception leads speakers to produce strong cues for prediction (i.e., substantial reduction of words with high vs. low discourse-dependent probability). Similarly, speakers who produce largest degrees of reduction also make the strongest predictions during perception, driving further perceptual learning. This set of results suggests a cyclic relationship between production and perception, with prediction driving perceptual learning for production and production driving further predictions (Dell & Chang, 2014).

Future research should continue to investigate this relationship between production and perception, in particular, the relationship between production and prediction, by focusing on specific processes that may be shared across modalities. Our focus in the current dissertation was on the process that generated phonetic variation from probabilistic information. A similar process generates prosodic variation (i.e., presence or absence of a pitch accent) from the same

probabilistic information. Similar to the results of Study 2 and the related studies that came before it (Arnold, 2008; Dahan, Tanenhaus, & Chambers, 2002), other studies have found that listeners use the congruent coupling of certain pitch accents (e.g., L + H*) and discourse status to make predictions (e.g., Ito & Speer, 2008).

Future work could investigate the relationship between L2 speakers' ability to produce the appropriate pitch accent for targets with high vs. low discourse-dependent probability and how they engage predictive processing during speech perception. Given that L2 speakers struggle to master other aspects of L2 prosody (e.g., perception of timing relations, Baese-Berk, Morrill, & Dilley, 2016; production of duration and F0, Aoyama & Guion, 2007), L2 speakers may also struggle to master the prosodic implementation of discourse status as well (e.g., Nava, 2008). However, if an L2 speaker exhibits mastery of this aspect of prosody in their productions, we would expect them to also use this probabilistic information to make predictions during speech perception (consistent with Hopp, 2013, 2016; Study 3). Furthermore, we would expect that L2 listeners who undergo perceptual training to learn to predict based on this information will learn to produce the pitch accent appropriate for targets with high vs. low discourse-dependent probability.

This hypothesis that perceptual learning driven by predictions contributes to production ability merits further investigation. In the General Discussion of Chapter 4, we discussed the possibility that with enough perceptual training, speakers who produce little reduction of high vs. low probability nouns could learn to produce larger levels of reduction. With enough training, we would predict that even an L1 speaker could learn new relations between probability and phonetic variation (e.g., low probability words should be reduced). Hopp (2016) demonstrated that L1 German listeners could un-learn the gender marking of German determiners via

perceptual learning of new statistics over the course of an experiment, suggesting that L1 listeners are willing to update their stable linguistic representations in the face of new information. Interesting questions arise when considering whether L1 speakers can learn from L2 speakers. If L2 speakers produce L1-like reduction of high vs. low probability content words (as our speakers in Study 1 do), perceptual training with L2 speech may lead to adaptation to new statistics by L1 listeners. If, as we discussed above, L1 listeners use discourse-dependent probabilistic reduction produced by L2 speakers as a predictive cue during speech perception, we would expect they can also learn from this speech. That is, L1 listeners who produce little reduction of nouns with high vs. low discourse-dependent probability may learn to do so via experience making predictions based on L2 phonetic variation. Answers to these questions would have implications for understanding communication between interlocutors with different levels of linguistic experience, and for understanding how such interactions impact the linguistic systems of the interlocutors over the course of a conversation.

5.4 Conclusions

The three studies of this dissertation provide a sketch of how discourse-dependent probabilistic information influences speech processing within and across production and perception. Throughout, we have considered how linguistic experience may impact these aspects of processing. Differences between individuals with different levels of linguistic experience emerged throughout these studies, allowing us to refine existing theories. As a whole, this dissertation has produced insights into – and has raised new questions about – one of the longest standing questions in linguistics research: the balance of variation and stability in the linguistic system.

REFERENCES

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439. http://doi.org/10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. http://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, *4*(11). doi:10.1371/journal.pone.0007678.

Aoyama, K., & Guion, S. G. (2007). Prosody in second language acquisition: Acoustic analyses of duration and F0 range. In O.-S. Bohn & M. Munro (Eds.), *Language experience in second language speech learning* (pp. 281–297). Amsterdam: John Benjamins Publishing Co.

Arnold, J. E. (2008). THE BACON not the bacon: How children and adults understand accented and unaccented noun phrases. *Cognition*, *108*, 69–99. doi:10.1016/j.cognition.2008.01.001

Arnold, J. E., Kahn, J. M., & Pancani, G. C. (2012). Audience design affects acoustic reduction via production facilitation. *Psychonomic Bulletin & Review*, *19*(3), 505–12. doi:10.3758/s13423-012-0233-y

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *33*(5), 914–930. http://doi.org/10.1037/0278-7393.33.5.914

Arnold, J. E., & Tanenhaus, M. K. (2007). Disfluency effects in comprehension: How new information can become accessible. In E. Gibson & N. Perlmutter (Eds.), *The processing and acquisition of reference* (pp. 197–217). MIT Press.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*(9), 578–582. doi:10.1111/j.0956-7976.2004.00723.x

Arnold, J. E., & Watson, D. G. (2015). Synthesising meaning and processing approaches to prosody: performance matters. *Language and Cognitive Processes, 30*, 88–102. doi:10.1080/01690965.2013.840733

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56. doi:10.1177/00238309040470010201

Baese-Berk, M. M., Morrill, T. H., & Dilley, L. C. (2016). Do non-native speakers use context speaking rate in spoken word Recognition? *Proceedings of the 8th International Conference on Speech Prosody (SP2016)*, 979–983.

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23–36. http://doi.org/10.1016/j.jml.2015.10.008

Baker, R. E., Baese-Berk, M., Bonnasse-Gahot, L., Kim, M., Van Engen, K. J., & Bradlow, A. R. (2011). Word durations in non-native English. *Journal of Phonetics*, *39*(1), 1–17. doi:10.1016/j.wocn.2010.10.006

Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, *52*(4), 391–413. doi:10.1177/0023830909336575

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42*, 1–22.

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. http://doi.org/10.1016/j.jml.2007.09.002

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*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. doi:10.1016/j.jml.2008.06.003

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English

conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024. doi:10.1121/1.1534836

Bock, K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 954-984). San Diego: CA: Academic Press.

Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: some consequences for comprehension. *Memory & Cognition*, *11*(1), 64–76. http://doi.org/10.3758/BF03197663

Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). Native "um"s elicit prediction of low-frequency referents , but non-native "um"s do not. *Journal of Memory and Language*, *75*, 104–116. http://doi.org/10.1016/j.jml.2014.05.004

Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.16, retrieved 6 April 2016 from http://www.praat.org/

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (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. http://doi.org/10.1121/1.418276

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS ONE*, *5*(5), 1-13.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency

measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

Bürki, A., Laganaro, M., & Alario, F. X. (2014). Phonologically driven variability: The case of determiners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1348–1362. http://doi.org/10.1037/a0036351

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68-86.

Caramazza, A., Miozzo, M., Costa, A., Schiller, N., & Alario, F.-X. (2001). A crosslinguistic investigation of determiner production. In E. Dupoux (Ed.), *Language, brain, and cognitive development: Essays in honor of Jacques Mehler* (pp. 209-226). Cambridge, MA: MIT Press.

Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1029–40. http://doi.org/10.1037/a0015901

Colomé, À. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of Memory and Language*, *45*(4), 721–736. doi:10.1006/jmla.2001.2793

Colomé, A., & Miozzo, M. (2010). Which words are activated during bilingual word production? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(1), 96–109. doi:10.1037/a0017677

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: implications for models of lexical access. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(5), 1283–96.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5–6), 507–534. http://doi.org/10.1080/01690960143000074

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465–480. http://doi.org/10.1006/jmla.1999.2688

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–67. http://doi.org/10.1006/cogp.2001.0750

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*, 292–314. http://doi.org/10.1016/S0749-596X(02)00001-3

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*(4), 313–349.

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*, 1-9. http://doi.org/10.1098/rstb.2012.0394

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. doi:10.1080/17470218.2012.720994

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2016). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, 1-14.

Dijsktra, T. & van Heuven, W. J. (1998). The BIA model and bilingual word recognition. *Localist connectionist approaches to human cognition*, 189-225.

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*(3), 294–311. http://doi.org/10.1016/j.jml.2008.06.006

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand. *Studies in Second Language Acquisition*, *35*(2), 353–387. http://doi.org/10.1017/S0272263112000915

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review*, *15*(4), 850–855. http://doi.org/10.3758/PBR.15.4.850

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language, 46*(1), 57-84.

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1–9. http://doi.org/10.1037/a0036756

Foucart, A., Ruiz-Tada, E., & Costa, A. (2015). How do you know I was about to say "book"? Anticipation processes affect speech processing and lexical recognition. *Language, Cognition & Neuroscience*, *30*(6), 768–780. http://doi.org/10.1080/23273798.2015.1016047

Fowler, C. A. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech*, *31*(4), 307–319. doi:10.1177/002383098803100401

Fowler, C. A, Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*(3), 396–413. http://doi.org/10.1016/S0749-596X(03)00072-X

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*, 489–504. doi:10.1016/0749-596X(87)90136-7

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2014). Reduction in prosodic prominence predicts speakers' recall: implications for theories of prosody. *Language, Cognition and Neuroscience*, *30*, 606–619. doi:10.1080/23273798.2014.966122

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, *80*(4), 748–775.

Garrett, M. F. (1975). The analysis of sentence production. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 133–175). San Diego: Academic Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–79.

Gollan, T. H., Montoya, R. I., Fennema-Notestine, C., & Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Memory & Cognition*, *33*(7), 1220–1234. doi:10.3758/BF03193224

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always a means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787–814. doi:10.1016/j.jml.2007.07.001

Gollan, T. H., Slattery, T. J., Goldenberg, D., Van Assche, E., Duyck, W., & Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *Journal of Experimental Psychology: General*, *140*(2), 186–209. doi:10.1037/a0022256

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, *28*, 191–215. http://doi.org/10.1177/0267658312437990

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. doi:10.1017/S0142716400002034

Gustafson, E., Engstler, C., & Goldrick, M. (2013). Phonetic processing of non-native speech in semantic vs non-semantic tasks. *The Journal of the Acoustical Society of America*, *134*(6), EL506. doi:10.1121/1.4826914

Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, *130*(4), 2139–2152. http://doi.org/10.1121/1.3623753

Hermans, D., Bongaerts, T., De Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, *1*(3), 213–229. doi:10.1017/S1366728998000364

Hernández, M., Costa, A., & Arnon, I. (2016). More than words: multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience*, *3798*(April), 1–16. doi:10.1080/23273798.2016.1152389

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56. http://doi.org/10.1177/0267658312461803

Hopp, H. (2015). Semantics and morphosyntax in predictive L2 sentence processing. *International Review of Applied Linguistics in Language Teaching*, *53*(3), 277–306. http://doi.org/10.1515/iral-2015-0014

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*. http://doi.org/10.1177/0267658315624960

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience, 31*(1), 19-31. http://doi.org/10.1080/23273798.2015.1072223

Isaacs, A. M., & Watson, D. G. (2010). Accent detection is a slippery slope: Direction and rate of F0 change drives listeners' comprehension. *Language and Cognitive Processes*, *25*(7–9), 1178–1200. http://doi.org/10.1080/01690961003783699

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*(2), 541–573. http://doi.org/10.1016/j.jml.2007.06.013

Jacobs, C. L., Yiu, L. K., Watson, D. G., & Dell, G. S. (2015). Why are repeated words produced with reduced durations? Evidence from inner speech and homophone production. *Journal of Memory and Language*, *84*, 37–48. doi:10.1016/j.jml.2015.05.004

Janssen, N., Schiller, N. O., & Alario, F.-X. (2014). The selection of closed-class elements during language production: A reassessment of the evidence and a new look on new data. *Language and Cognitive Processes*, *29*(6), 695–708. http://doi.org/10.1080/01690965.2012.693617

Jescheniak, J. D., Schriefers, H., & Lemhöfer, K. (2014). Selection of freestanding and bound gender-marking morphemes in speech production: a review. *Language, Cognition and Neuroscience*, *29*(6), 684–694. http://doi.org/10.1080/01690965.2012.654645

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: Benjamin.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*, 257–282. http://doi.org/10.1075/lab.4.2.05kaa

Kaan, E. (2016). Susceptibility to interference: underlying mechanisms, and implications for prediction. *Bilingualism: Language and Cognition*, 1–2. http://doi.org/10.1017/S1366728916000894

Kahn, J. M., & Arnold, J. E. (2012). A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language*, *67*(3), 311–325. doi:10.1016/j.jml.2012.07.002

Kahn, J. M., & Arnold, J. E. (2015). Articulatory and lexical repetition effects on durational reduction: speaker experience vs. common ground. *Language, Cognition and Neuroscience, 30*(1-2), 103-119. doi:10.1080/01690965.2013.848989

Kittredge, A. K., & Dell, G. S. (2016). Learning to speak by listening: Transfer of phonotactics from perception to production. *Journal of Memory and Language*, *89*, 8–22. http://doi.org/10.1016/j.jml.2015.08.001

Kleinschmidt, D., & Jaeger, T. F. (2016). What do you expect from an unfamiliar talker? Abstract from *LabPhon15*. Ithaca, NY.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition & Neuroscience*, *31*(1), 32–59. http://doi.org/10.1080/23273798.2015.1102299

Ladd, R. (1996) *Intonational phonology*. Cambridge: University Press.

Lam, T. Q., & Marian, V. (2015). Repetition reduction during word and concept overlap in bilinguals. *Journal of Memory and Language*, *84*, 88–107. doi:10.1016/j.jml.2015.05.005

Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: why repeated referents have reduced prominence. *Memory & Cognition*, *38*(8), 1137-1146. doi:10.3758/MC.38.8.1137

Lam, T. Q., & Watson, D. G. (2014). Repetition reduction: Lexical repetition in the absence of referent repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 829–843.

Lapointe, S. G., & Dell, G. S. (1989). A synthesis of some recent work in sentence production. In G. N. Carlson and M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 107–156). Spring Netherlands.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. doi:10.3758/s13428-011-0146-0

Levelt, W. J., Roelofs, A., & Meyer, a S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS) 19* (pp. 849–856). Cambridge, MA: MIT Press.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447–464. doi:10.1016/j.jml.2010.07.003

Liberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*(3), 172–187.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439).

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. http://doi.org/10.1037/a0029284

Martin, C. D., Thierry, G., Kuipers, J., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, *69*(4), 574–588. http://doi.org/10.1016/j.jml.2013.08.001

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. http://doi.org/10.1016/0010-0285(86)90015-0

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), 33–42. http://doi.org/10.1016/S0010-0277(02)00157-9

Meagher, B. R., & Fowler, C. A. (2014). Embedded articulation: shifts in location influence speech production. *Language, Cognition and Neuroscience, 29*(5), 1–7.

Meyer, A. S., & Damian, M. F. (2007). Activation of distractor names in the picture-picture interference paradigm. *Memory & Cognition*, *35*(3), 494–503.

Miozzo, M., & Caramazza, A. (1999). The selection of determiners in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 907–922. http://doi.org/10.1037/0278-7393.25.4.907

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R*. Chapman and Hall/CRC.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. http://doi.org/10.1016/j.jml.2007.11.006

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: evidence from the shadowing task. *Cognition*, *109*(1), 168–73. http://doi.org/10.1016/j.cognition.2008.08.002

Mitterer, H., & McQueen, J. M. (2009). Processing reduced word-forms in speech perception using probabilistic knowledge about speech production. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 244–63. http://doi.org/10.1037/a0012730

Mitterer, H., & Russell, K. (2013). How phonological reductions sometimes help the listener. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 977–84. doi:10.1037/a0029196

Morales, L., Paolieri, D., Dussias, P. E., Valdés Kroff, J. R., Gerfen, C., & Bajo, M. T. (2016). The gender congruency effect during bilingual spoken-word recognition. *Bilingualism: Language and Cognition, 19*(2), 294-310.

Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 555–563. doi:10.1037//0278-7393.28.3.555

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Munson, B., & Solomon, N. P. (2004). The Effect of Phonological Neighborhood Density on Vowel Articulation. *Journal of Speech, Language, and Hearing Research*, *47*, 1048–1058.

Nava, E. (2008). Prosody in L2 acquisition. *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference*. Somerville, MA: Cascadilla Proceedings Project, pp. 155–164.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. http://doi.org/10.1016/j.wocn.2010.12.007

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–95. http://doi.org/10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience, 31*(1), 4–18. http://doi.org/10.1080/23273798.2015.1081703

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17. doi:10.1016/j.jml.2014.10.003

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–47. http://doi.org/10.1017/S0140525X12001495

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. L. Bybee & P. J. Hopper (Eds.), *Frequency effects and emergent grammar* (pp. 137–157). Amsterdam: John Benjamins Publishing Co.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 101–139). Berlin/New York: Mouton.

Pluymaekers, M., Ernestus, M., Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica, 62*(2-4), 149-159.

Puckette, M., Zicarelli, D., Sussman, R., Clayton, J. K., Bernstein, J., Nevile, B., Place, T., Grosse, D., Dudas, R., Jourdan, E., Lee, M., & Schabtach, A. (2011). MAX/MSP [Computer program]. Version 5.1.9 retrieved from https://cycling74.com/downloads/older/#.V7UrUWUqafQ.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing [Computer program]. Version 3.2.4 retrieved from https://www.R-project.org/.

Rohde, H., & Ettlinger, M. (2012). Integration of pragmatic and phonetic cues in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 967–983.

Rosa, E. C., Finch, K. H., Bergeson, M., & Arnold, J. E. (2015). The effects of addressee attention on prosodic prominence. *Language, Cognition and Neuroscience*, *30*(1-2), 48–56. doi:10.1080/01690965.2013.772213

Runnqvist, E., Gollan, T. H., Costa, A., & Ferreira, V. S. (2013). A disadvantage in bilingual sentence production modulated by syntactic frequency and similarity across languages. *Cognition*, *129*(2), 256–63. doi:10.1016/j.cognition.2013.07.008

Sadat, J., Martin, C. D., Alario, F. X., & Costa, A. (2012). Characterizing the bilingual disadvantage in noun phrase production. *Journal of Psycholinguistic Research*, *41*(3), 159–79. doi:10.1007/s10936-011-9183-1

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically-modulated sub-phonetic variation on lexical competition. *Cognition*, *105*(2), 466–476.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*, 145–163. http://doi.org/10.1016/j.jml.2013.11.002

Scarborough, R. A. (2010). Lexical and contextual predictability: Confluent effects on the production of vowels. In C. Fougeron, M. Kuhnert, & M. D'Imperio (Eds.), *Laboratory Phonology 10* (pp. 557–586). Berlin: De Gruyter.

Schertz, J., & Ernestus, M. (2014). Variability in the pronunciation of non-native English the: Effects of frequency and disfluencies. *Corpus Linguistics and Linguistic Theory*, *10*(2), 329–345. doi:10.1515/cllt-2014-0024

Selkirk, E. (1996). The prosodic structure of function words. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstraping from speech to grammar in early acquisition* (pp. 187–213). Mahwah, NJ: Lawrence Erlbaum.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155. doi:10.1016/j.cognition.2014.06.013

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language*, *25*(1), 169–201. doi:10.1017/S0305000997003395

Shook, A., Goldrick, M., Engstler, C., & Marian, V. (2014). Bilinguals show weaker lexical access during spoken sentence comprehension. *Journal of Psycholinguistic Research*. http://doi.org/10.1007/s10936-014-9322-6

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033. http://doi.org/10.1121/1.1531176

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Appendix A

Stimuli

Table A1. Pairs of target items. * denotes items from BOSS. Frequencies are from the SUBTLEX corpus (occurrences per million words).

Target item 1	L1 name agreement	L2 name agreement	Frequency	Target item 2	L1 name agreement	L2 name agreement	Frequency
arch	78.9%	70%	3.7	arm*	94.7%	100%	65.4
baby	100%	80%	509.4	bacon	100%	80%	11.9
bag	100%	70%	94.0	bat*	100%	60%	20.6
beach	100%	90%	56.6	bee*	94.7%	100%	10.4
bed*	100%	100%	187.1	bench*	78.9%	60%	9.7
brick*	94.7%	70%	10.2	bridge*	100%	100%	45.7
butter	94.7%	60%	20.4	button*	100%	40%	28.3
cage	63.2%	60%	20.3	cake	100%	80%	45.1
camel*	100%	90%	5.0	camera*	100%	100%	57.0
candy*	89.5%	100%	35.8	candle*	100%	100%	8.0
chair*	100%	100%	49.2	chain*	100%	70%	21.2
dollar	89.5%	50%	27.7	dolphin*	94.7%	100%	2.8
fish	84.2%	100%	83.5	fist	100%	40%	7.4
ghost	100%	80%	36.6	goat	100%	70%	10.5
mouse*	78.9%	70%	19.1	mouth	100%	90%	104.4
net	100%	60%	15.5	nest*	100%	100%	59.5
peanuts	100%	100%	12.4	pizza	100%	100%	33.5
pickle*	100%	60%	4.6	picture	78.9%	60%	138.5
pig*	94.7%	100%	39.1	pill*	89.5%	50%	11.8
sandwich	100%	90%	21.9	sandal*	89.5%	40%	0.2
scar	63.2%	60%	8.5	star	100%	90%	81.4
tire	100%	80%	12.4	tie*	94.7%	90%	44.4
turtle	100%	90%	17.0	turkey	100%	40%	22.6
witch	100%	50%	27.6	wing	94.7%	100%	20.2

Table A2. Non-target items with associated target pairs. * denotes items from BOSS.

Target pair	Non-target item 1	Non-target item 2	Non-target item 3
arch/arm	pepper*	bottle	clock*
baby/bacon	leaf*	heart	shirt*
bag/bat	table*	onion*	daisy*
beach/bee	eye*	watch*	eggs*
bed/bench	headphones*	sock*	wine
brick/bridge	hand*	airplane*	phone
butter/button	monkey	hand*	dog
cage/cake	lemon*	moon*	belt*
camel/camera	lock*	apple	bird*
candy/candle	ear*	bell	toothbrush*
chair/chain	box	window	coins*
dollar/dolphin	grapes	nose*	fork*
fish/fist	key*	road*	guitar*
ghost/goat	cat	sun	bra*
mouse/mouth	basket*	car	spoon*
net/neck	tree*	snake*	cookie*
peanuts/pizza	ruler*	glasses	comb*
pickle/picture	horse*	shoe*	ring
pig/pill	orange*	cart	skirt
sandwich/sandal	book*	coat	chicken*
scar/star	door	pen*	tank*
tire/tie	kite*	knife*	spider
turtle/turkey	hat*	bowl	scarf*
witch/wing	brain	pencil*	toilet*

Appendix B

Measurement Criteria for Acoustic Analyses of Study 1

Table A3. Measurement criteria for marking noun onset boundaries according to initial phoneme/class of the noun.

Phoneme/Class	Criteria
Vowels	F2 lowering Pause between determiner and noun
[w]	Amplitude drop off in spectrogram Before lowering of F2 and raising of F3
Stops/affricates	Before release burst
Fricatives	Onset of high energy noise in spectrogram Noise evident in waveform
Nasals	Overall drop in amplitude in spectrogram Shape change in waveform

Table A4. Measurement criteria for marking noun offset boundaries for [f]-initial verb according to final phoneme/class of the noun.

Phoneme/Class	Criteria
Sonorants	Onset of high frequency energy in spectrogram and/or waveform
Stops	End of aspiration Lower overall amplitude in energy bands at lowest frequencies
[ɵ]	Drop in overall amplitude

Table A5. Measurement criteria for marking noun offset boundaries for [ɛ]-initial verb according to final phoneme/class of the noun.

Phoneme/Class	Criteria
[ɹ]	After raising of F2 and F3 Onset of stable F3
[n]	Sudden amplitude increase Increased clarity of F2 formant structure
[l]	F2 rising
[aɪ]	Stability of formants after diphthong movement
[ɪ]	Before lowering of F2 and F3
Fricatives/affricates	Decreasing amplitude of noise in spectrogram Onset of periodicity
Stops	End of aspiration Onset of periodicity

Table A6. Measurement criteria for marking noun offset boundaries for [ʃ]-initial verb according to final phoneme/class of the noun.

Phoneme/Class	Criteria
Sonorants	Onset high frequency energy in spectrogram and waveform
Stops	End of aspiration Onset of periodicity

Table A7. Measurement criteria for marking noun offset boundaries for [ɹ]-initial verb according to final phoneme/class of the noun.

Phoneme/Class	Criteria
[n]	After lowering of F3
Other sonorants	Onset of decrease in overall amplitude After lowering of F3 to F2

Appendix C

Model Output for Control Factors in Study 1

Table A8. Results for control factors in model of response times.

	Beta	SE	Chi-squared	*p*
Block	0.10	0.01	32.20	< 0.001
Log determiner duration	0.07	0.01	23.26	< 0.001
Log noun duration	0.07	0.01	21.23	< 0.001

Table A9. Results for control factors in model of noun durations.

	Beta	SE	Chi-squared	*p*
Block	-0.02	0.01	2.08	> 0.05
Log frequency	-0.05	0.02	5.35	< 0.05
Log RT	0.13	0.03	15.28	< 0.001

Table A10. Results for control factors in model of determiner durations.

	Beta	SE	Chi-squared	*p*
Block	0.03	0.02	3.00	< 0.09
Log RT	0.23	0.04	25.93	< 0.001
Log noun duration	0.32	0.04	47.39	< 0.001

Appendix D

Model Output for Looks to the Target in the Late Window in Experiment 1 of Study 2

	Estimate	SE	Chi-squared	*p*
Intercept	-0.16	0.12		
ot1	2.26	0.60	12.54	0.0004***
ot2	-4.89	0.38	72.12	2.03e-17***
ot3	0.65	0.30	4.48	0.034 *
ot4	0.71	0.17	14.00	0.0001 ***
Reduce	0.12	0.06	4.43	0.035 *
Given	-0.99	0.06	311.00	8.03e-70 ***
Group	0.36	0.25	2.07	0.151
Reduce:Given	-0.07	0.11	0.38	0.537
Reduce:Group	-0.41	0.11	13.81	0.0002 ***
Given:Group	0.28	0.11	6.53	0.011 *
Reduce:Given:Group	0.30	0.22	1.88	0.171
ot1:Reduce	-2.35	0.28	68.85	1.06e-16 ***
ot2:Reduce	-1.33	0.28	22.29	2.34e-06 ***
ot3:Reduce	0.84	0.28	8.91	0.003 **
ot4:Reduce	0.53	0.28	3.46	0.063
ot1:Given	-0.62	0.28	4.87	0.028 *
ot2:Given	0.42	0.28	2.23	0.135
ot3:Given	2.69	0.28	89.59	2.92e-21 ***
ot4:Given	-1.33	0.28	22.25	2.40e-06 ***
ot1:Group	-2.80	1.20	5.19	0.023 *
ot2:Group	-1.47	0.76	3.62	0.057
ot3:Group	0.71	0.60	1.38	0.240
ot4:Group	0.81	0.34	5.46	0.020 *
ot1:Reduction:Given	0.12	0.57	0.04	0.833
ot2:Reduction:Given	-1.21	0.57	4.55	0.033 *
ot3:Reduction:Given	0.08	0.57	0.02	0.887
ot4:Reduction:Given	-0.004	0.57	0	0.994
ot1:Reduction:Group	0.59	0.57	1.08	0.299
ot2:Reduction:Group	1.20	0.57	4.54	0.033 *
ot3:Reduction:Group	-0.01	0.57	0.0002	0.987
ot4:Reduction:Group	1.17	0.57	4.32	0.038 *
ot1:Given:Group	-1.33	0.57	5.51	0.019 *
ot2:Given:Group	0.59	0.57	1.08	0.299
ot3:Given:Group	0.15	0.57	0.07	0.785
ot4:Given:Group	-0.29	0.57	0.26	0.612
ot1:Reduction:Given:Group	6.35	1.13	31.44	2.05e-08 ***
ot2:Reduction:Given:Group	4.63	1.13	16.76	4.25e-05 ***
ot3:Reduction:Given:Group	-1.41	1.13	1.56	0.211
ot4:Reduction:Given:Group	-0.29	1.13	0.07	0.796

Appendix E

Model Output for Looks to the Competitor in the Late Window in Experiment 1 of Study 2

	Estimate	SE	Chi-squared	*p*
Intercept	-4.27	0.07		
ot1	-5.22	0.31	91.71	1.01e-21 ***
ot2	2.44	0.24	54.65	1.44e-13 ***
ot3	0.66	0.24	7.11	0.008 **
ot4	-1.51	0.20	38.25	6.22e-10 ***
Reduce	-0.35	0.05	41.69	1.07e-10 ***
Given	0.59	0.05	117.28	2.49e-27 ***
Group	-0.08	0.14	0.37	0.544
Reduce:Given	-0.99	0.11	83.77	5.56e-20 ***
Reduce:Group	0.52	0.11	22.97	1.64e-06 ***
Given:Group	-0.04	0.11	0.17	0.682
Reduce:Given:Group	0.26	0.22	1.42	0.233
ot1:Reduce	0.33	0.28	1.46	0.227
ot2:Reduce	1.06	0.28	14.90	0.0001 ***
ot3:Reduce	-0.32	0.28	1.39	0.238
ot4:Reduce	-0.51	0.28	3.40	0.065
ot1:Given	-0.06	0.28	0.05	0.819
ot2:Given	0.39	0.28	2.01	0.157
ot3:Given	-1.28	0.28	21.64	3.29e-06 ***
ot4:Given	1.32	0.28	22.79	1.81e-06 ***
ot1:Group	-0.01	0.63	0	0.979
ot2:Group	1.05	0.48	4.49	0.034 *
ot3:Group	-0.60	0.48	1.54	0.214
ot4:Group	-1.24	0.40	8.96	0.003 **
ot1:Reduction:Given	2.45	0.55	19.84	8.42e-06 ***
ot2:Reduction:Given	0.50	0.55	0.84	0.360
ot3:Reduction:Given	-2.57	0.55	21.72	3.15e-06 ***
ot4:Reduction:Given	0.06	0.55	0.01	0.912
ot1:Reduction:Group	-0.71	0.55	1.67	0.197
ot2:Reduction:Group	-0.50	0.55	0.82	0.364
ot3:Reduction:Group	-0.16	0.55	0.08	0.771
ot4:Reduction:Group	-0.32	0.55	0.34	0.559
ot1:Given:Group	3.22	0.55	34.15	5.09e-09 ***
ot2:Given:Group	-0.53	0.55	0.92	0.336
ot3:Given:Group	-0.33	0.55	0.36	0.551
ot4:Given:Group	0.003	0.55	0	0.996
ot1:Reduction:Given:Group	-1.31	1.10	1.42	0.233
ot2:Reduction:Given:Group	0.85	1.10	0.60	0.439
ot3:Reduction:Given:Group	2.34	1.10	4.51	0.034 *
ot4:Reduction:Given:Group	-1.89	1.10	2.95	0.086

Appendix F

Model Output for Looks to the Target in the Late Window in Experiment 2 of Study 2

	Estimate	SE	Chi-squared	*p*
Intercept	-0.15	0.16		
ot1	1.98	0.82	5.27	0.022 *
ot2	-5.03	0.66	29.28	6.25e-08 ***
ot3	-0.31	0.41	0.56	0.455
ot4	-0.27	0.28	0.92	0.337
Reduce	0.16	0.17	0.89	0.346
Given	-1.03	0.24	14.05	0.0002 ***
Reduce:Given	-0.10	-0.10	0.09	0.762
ot1:Reduce	-1.29	0.36	12.77	0.0004 ***
ot2:Reduce	-2.13	0.36	34.81	3.64e-09 ***
ot3:Reduce	1.13	0.36	9.76	0.002 **
ot4:Reduce	-0.96	0.36	7.04	0.008 **
ot1:Given	-1.35	0.36	13.91	0.0002 ***
ot2:Given	1.61	0.36	19.81	8.58e-06 ***
ot3:Given	2.58	0.36	50.88	9.84e-13 ***
ot4:Given	-1.39	0.36	14.80	0.0001 ***
ot1:Reduction:Given	-0.07	0.72	0.01	0.917
ot2:Reduction:Given	-1.12	0.72	2.40	0.121
ot3:Reduction:Given	-0.53	0.72	0.55	0.460
ot4:Reduction:Given	-0.02	0.72	0.001	0.972

Appendix G

Model Output for Looks to the Competitor in the Late Window in Experiment 2 of Study 2

	Estimate	SE	Chi-squared	*p*
Intercept	-4.44	0.10		
ot1	-4.85	0.51	37.60	8.67e-10 ***
ot2	1.88	0.44	13.65	0.0002 ***
ot3	0.44	0.25	2.98	0.084
ot4	-0.79	0.28	6.82	0.009 **
Reduce	-0.32	0.22	2.01	0.1567
Given	0.83	0.24	9.92	0.002 **
Reduce:Given	-1.25	0.39	8.61	0.003 **
ot1:Reduce	-0.01	0.99	0.0002	0.989
ot2:Reduce	1.89	0.86	4.37	0.036 *
ot3:Reduce	-0.92	0.70	1.68	0.195
ot4:Reduce	-0.29	0.48	0.36	0.548
ot1:Given	-0.86	1.26	0.46	0.497
ot2:Given	-2.00	0.78	5.86	0.016 *
ot3:Given	-0.10	0.76	0.02	0.899
ot4:Given	1.16	0.52	4.61	0.032 *
ot1:Reduction:Given	2.15	1.48	2.02	0.155
ot2:Reduction:Given	2.10	1.25	8.93	0.003 **
ot3:Reduction:Given	-2.65	1.16	4.70	0.030 *
ot4:Reduction:Given	-1.10	1.03	1.11	0.292

Appendix H

Output of Production Model in Study 3

	Estimate	SE	Chi-squared	*p*
Intercept	-0.93	0.04		
Block	-0.02	0.01	1.26	0.262
Response time	0.15	0.05	7.47	0.006
Frequency	-0.05	0.02	6.38	0.012*
Discourse condition	0.08	0.01	27.16	1.87e-07***
Quadratic target BLUPs	-0.003	0.008	0.15	0.694
Intercept competitor BLUPs	0.02	0.02	1.53	0.216
Cubic competitor BLUPs	-0.02	0.008	5.83	0.016*
Discourse condition:quadratic target BLUPs	-0.001	0.003	0.002	0.967
Discourse condition:cubic competitor BLUPs	-0.009	0.007	1.80	0.180
Discourse condition:overall competitor BLUPs	0.007	0.003	5.66	0.017*

Appendix I

Output of Perception Model for Looks to the Target in Study 3

	Estimate	SE	Chi-squared	*p*
Intercept	-0.34	0.17		
ot1	0.72	0.17	13.20	0.0003***
ot2	-0.82	0.11	29.47	5.67e-08***
ot3	0.06	0.09	0.42	0.517
ot4	0.06	0.04	1.85	0.173
Reduce	0.16	0.11	1.99	0.159
Given	0.57	0.13	14.35	0.0002***
BLUPs	-0.13	0.18	0.56	0.456
Reduce:Given	0.06	0.04	2.22	0.137
Reduce:BLUPs	-0.04	0.11	0.10	0.750
Given:BLUPs	0.07	0.13	0.33	0.563
Reduce:Given:BLUPs	0.10	0.04	7.77	0.005**
ot1:Reduce	-0.26	0.04	48.70	2.20e-12***
ot2:Reduce	-0.19	0.04	26.23	3.03e-07***
ot3:Reduce	0.08	0.04	5.05	0.025*
ot4:Reduce	-0.006	0.04	0.03	0.869
ot1:Given	-0.004	0.04	0.01	0.917
ot2:Given	-0.01	0.04	0.12	0.733
ot3:Given	-0.26	0.04	47.37	5.88e-12***
ot4:Given	0.12	0.04	9.94	0.002**
ot1:BLUPs	-0.35	0.16	4.23	0.040*
ot2:BLUPs	-0.23	0.10	4.88	0.027*
ot3:BLUPs	0.10	0.09	1.10	0.294
ot4:BLUPs	0.01	0.04	0.11	0.744
ot1:Reduction:Given	0.15	0.04	16.33	5.31e-05***
ot2:Reduction:Given	0.17	0.04	21.67	3.24e-06***
ot3:Reduction:Given	-0.04	0.04	1.09	0.297
ot4:Reduction:Given	-0.007	0.04	0.04	0.851
ot1:Reduction:BLUPs	-0.10	0.04	7.70	0.006**
ot2:Reduction:BLUPs	-0.07	0.04	3.13	0.077
ot3:Reduction:BLUPs	-0.07	0.04	3.29	0.070
ot4:Reduction:BLUPs	0.02	0.04	0.28	0.595
ot1:Given:BLUPs	0.07	0.04	3.93	0.047*
ot2:Given:BLUPs	0.07	0.04	3.49	0.062
ot3:Given:BLUPs	-0.01	0.04	0.09	0.762
ot4:Given:BLUPs	0.05	0.04	1.69	0.194
ot1:Reduction:Given:BLUPs	0.19	0.04	26.57	2.54e-07***
ot2:Reduction:Given:BLUPs	-0.0004	0.04	0.0001	0.991
ot3:Reduction:Given:BLUPs	-0.10	0.04	7.12	0.008**
ot4:Reduction:Given:BLUPs	-0.07	0.04	3.17	0.075

Appendix J

Output of Perception Models for Looks to the Competitor in Study 3

	Estimate	SE	Chi-squared	*p*
Intercept	-4.22	0.08		
ot1	-1.02	0.09	42.39	7.46e-11***
ot2	0.38	0.06	24.18	8.80e-07***
ot3	0.19	0.06	7.81	0.005**
ot4	-0.17	0.05	11.44	0.0007***
Reduce	-0.30	0.04	66.72	3.13e-16***
Given	-0.31	0.04	67.15	2.52e-16***
BLUPs	0.08	0.08	0.97	0.325
Reduce:Given	0.28	0.04	56.78	4.87e-14***
Reduce:BLUPs	-0.27	0.04	53.36	2.78e-13***
Given:BLUPs	0.20	0.04	29.86	4.63e-08***
Reduce:Given:BLUPs	-0.03	0.04	0.74	0.389
ot1:Reduce	0.07	0.04	3.33	0.068
ot2:Reduce	0.13	0.04	12.09	0.0005***
ot3:Reduce	-0.02	0.04	0.42	0.516
ot4:Reduce	-0.03	0.04	0.85	0.358
ot1:Given	0.16	0.04	19.64	9.37e-06***
ot2:Given	-0.06	0.04	3.01	0.083
ot3:Given	0.11	0.04	8.77	0.003**
ot4:Given	-0.13	0.04	12.11	0.0005***
ot1:BLUPs	0.07	0.10	0.56	0.456
ot2:BLUPs	-0.01	0.06	0.03	0.874
ot3:BLUPs	0.003	0.06	0.03	0.960
ot4:BLUPs	0.06	0.05	1.42	0.234
ot1:Reduction:Given	-0.15	0.04	16.96	3.82e-05***
ot2:Reduction:Given	-0.004	0.04	0.01	0.918
ot3:Reduction:Given	0.18	0.04	24.44	7.67e-07***
ot4:Reduction:Given	-0.05	0.04	1.78	0.182
ot1:Reduction:BLUPs	0.25	0.04	45.38	1.62e-11***
ot2:Reduction:BLUPs	0.07	0.04	4.03	0.045*
ot3:Reduction:BLUPs	-0.04	0.04	1.20	0.273
ot4:Reduction:BLUPs	0.07	0.04	4.08	0.044*
ot1:Given:BLUPs	0.00006	0.04	0.003	0.986
ot2:Given:BLUPs	0.02	0.04	0.31	0.576
ot3:Given:BLUPs	0.03	0.04	0.66	0.418
ot4:Given:BLUPs	-0.04	0.04	1.01	0.316
ot1:Reduction:Given:BLUPs	-0.14	0.04	14.59	0.0001***
ot2:Reduction:Given:BLUPs	-0.13	0.04	12.00	0.0005***
ot3:Reduction:Given:BLUPs	0.13	0.04	13.11	0.0003***
ot4:Reduction:Given:BLUPs	0.001	0.04	0.001	0.972