Assessing the listener-oriented account of predictability-based phonetic reduction

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy in the Graduate School of The Ohio State University

By

Rory John Turnbull, MA (Hons), MA

Graduate Program in Linguistics

The Ohio State University
2015

Dissertation Committee:

Cynthia G. Clopper, Advisor
Mary E. Beckman
Rebecca Morley
Shari R. Speer
Abstract

Phonetic reduction is a common feature of everyday speech. Numerous studies have documented that words, syllables, and other linguistic elements which are more predictable are pronounced with less acoustic prominence than words, syllables, and elements which are less predictable. This phenomenon is referred to as predictability-based phonetic reduction.

Several accounts of this phenomenon exist. The focus of this dissertation is the listener-oriented account, which theorizes that predictability-based phonetic reduction arises from an interaction between the competing forces of conservation of effort and conservation of intelligibility. From this perspective, talkers use the least effort possible to provide the maximum level of comprehension—that is, their speech productions are guided by consideration of the listener's needs. For elements which are predictable, likelihood of comprehension is relatively high, and thus the talker is free to conserve effort and produce the element in a phonetically reduced way. For unpredictable elements, however, likelihood of comprehension is relatively low, and thus the talker must produce the element in a clear manner.

This dissertation presents the results of nine experiments examining different aspects of the listener-oriented account. The first three experiments tested the prediction of this account that individual theory of mind ability is positively correlated with extent of phonetic reduction. Results suggest that no such relationship exists for the
variables of lexical frequency, phonological neighborhood density, and second mention reduction. For semantic predictability, however, a negative correlation was observed, such that talkers with poor theory of mind ability had a greater extent of phonetic reduction than talkers with good theory of mind ability. These results fail to support the listener-oriented prediction.

The next three experiments tested the foundational assumption of the listener-oriented account that unreduced speech is easier for the listener to process than reduced speech. This assumption was tested at multiple levels of processing: subjective judgements of speech clarity, speech intelligibility, lexical decision, and semantic acceptability. Results suggest that, at all of these levels of processing, unreduced speech facilitates lexical retrieval relative to reduced speech. This finding is consistent with the assumption of the listener-oriented account. However, significant and systematic individual variation in responses was observed to be partially modulated by theory of mind ability. This unexpected finding is not predicted under a listener-oriented account.

The final three experiments used phonetic corpus analysis to investigate the extent to which predictability-based reduction can be attributed to a single factor, such as listener orientation. Results suggest that at least two and probably three factors are required to adequately model these effects: lexical and contextual factors are distinct, and contextual factors could be further split into discourse-specific and domain-general factors. This result is not consistent with major theory of predictability-based reduction, but can be accounted for under a hybrid model, combining an egocentric model of common ground with an exemplar-dynamic model of the lexicon.

Taken together, these results suggest that the listener-oriented account enjoys limited explanatory adequacy. The results and implications for this study are discussed in terms of our understanding of speech production, individual differences, and speech communication.
for Talia
Acknowledgements

If getting a PhD is a lot like getting married,\(^2\) then writing the acknowledgements section is like creating the guest list for a wedding. There is a lot of freedom to have an extravagant ceremony the whole town is invited to, or a low-key event at the registrar’s office. There is a strict implicational hierarchy of who gets invited. There is lots of swithering, contemplation, and head-scratching. Regardless, everyone is happy that you have come this far and wishes you well as you begin the next phase of life.

First and foremost, none of this work would have been possible without the guidance, support, and encouragement of my advisor, Cynthia Clopper. Her advising and mentoring, both direct and indirect, have been instrumental in not only the development of the research reported herein, but in my own personal development as a researcher with an open mind. Cynthia embodies the attributes of an outstanding teacher, mentor, scholar, and colleague, a model which I can only hope to emulate in my own career. My dissertation committee has also proved vital in helping me to see this project to completion. All have provided a listening ear and careful feedback, and demonstrated a remarkable degree of patience and forbearance of my faltering

\(^1\)I apologize in advance to all who I have inadvertently omitted from these acknowledgements. Please be assured that your absence is likely due to my own carelessness and forgetfulness, and not a reflection of the value of your contribution.

\(^2\)http://www.phdcomics.com/comics/archive.php?comicid=1296
attempts to become a better scientist. Mary Beckman has helped me to consider the broader implications of any results and encouraged open-mindedness in my interpretations. Becca Morley was always willing to remind me of the importance of stepping through the logic of every claim and never taking anything for granted. Shari Speer has been a source of constant support and guidance, particularly when reflecting on unexpected results.

This dissertation was partially funded by Ohio State University’s Presidential Fellowship, which I gratefully acknowledge. I would also like to thank those who helped me with my fellowship application, including my dissertation committee, Matt Goldrick, Cindy Johnson, Brian Joseph, Bridget Smith, and Alan Yu. I am grateful to my undergraduate research assistants Rachel Monnin, McKenna Reeher, and Amy Roberts for their assistance in data collection and coding. Greg Kierstead, Liz McCullough, and Mike Phelan all offered their vocal talent for the construction of various stimuli, and I thank them for their willingness to help.

My officemates, Lara Downing, Jeff Holliday, Sara Phillips-Bourass, Liz McCullough, Andy Plummer, and Pat Reidy, have been a source of support, both intellectual and moral, for all of my years in graduate school. Despite the occasional hardship, graduate school was a joyful experience, largely due to past and present OSU graduate students with whom I have built bonds of friendship, including Sonia Barnes, Alec Buchner, Rachel Steindel Burdin, Whitney Chappell, Dave Howcroft, Evan Jaffe, Cindy Johnson, Greg Kierstead, Mike Phelan, Tsz-Him Tsui, Marty van Schijndel, Bridget Smith, Abby Walker, Kodi Weatherholtz, and Seth Wiener. The Bahá’í community of central Ohio has been a tremendous source of energy, inspiration, and constancy, and I am deeply moved by the faithfulness and love manifested by this committed group of believers. Likewise, OSU’s Bahá’í Campus Association has provided me with a spiritual
anchor point and a frequent reminder that the well-being of humankind is the ultimate aim of any worthwhile endeavor.

Great gratitude is also due to Kathryn Campbell-Kibler for helping me navigate the labyrinthine depths of the OSU IRB, and Brett Gregory, Julie McGory, Claudia Morettini, and Julia Papke for helping me navigate other equally byzantine aspects of OSU’s administration. I would not have begun this journey were it not for the encouragement and mentorship of Bob Ladd and Peter Graff, who helped me to see where I wanted to go and how to get there. I hope that I have made them proud. I also wish to thank everyone who has challenged me, who has proved me wrong, and who has caused me to keep an open and inquiring mind.

My parents, John and Fiona, deserve special thanks for their support and encouragement at every stage. To them I owe a great deal, too much to enumerate in words. I am also grateful to my parents-in-law Ed and Marilyn, for providing me with a home away from home and for their encouragement in the final stages of the writing process. Finally, I am indebted to Talia, for her patience and endearing love throughout the process. I would be blessed to have spent but a single day with her.
Vita

2009 .......................... MA (Hons) English Language & Linguistics, University of Edinburgh

2013 .......................... MA Linguistics, The Ohio State University

Publications


Turnbull, R. and Clopper, C. G. (2013). Effects of semantic predictability and dialect variation on vowel production in clear and plain lab speech. *Proceedings of Meetings on*

Fields of Study

Major Field: Linguistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xvi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xviii</td>
</tr>
<tr>
<td><strong>1 Framing the problem</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Predictability effects in speech production</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Observed predictability effects and their properties</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1.1 Lexical frequency</td>
<td>3</td>
</tr>
<tr>
<td>1.1.1.2 Phonological neighborhood density</td>
<td>4</td>
</tr>
<tr>
<td>1.1.1.3 Discourse mention</td>
<td>8</td>
</tr>
<tr>
<td>1.1.1.4 Semantic predictability</td>
<td>9</td>
</tr>
<tr>
<td>1.1.1.5 Other predictability effects</td>
<td>10</td>
</tr>
<tr>
<td>1.1.2 The source of predictability effects</td>
<td>11</td>
</tr>
<tr>
<td>1.2 Theoretical perspectives</td>
<td>12</td>
</tr>
<tr>
<td>1.2.1 The listener-oriented perspective</td>
<td>12</td>
</tr>
<tr>
<td>1.2.1.1 Evidence for listener-orientation</td>
<td>14</td>
</tr>
<tr>
<td>1.2.1.2 Smooth signal redundancy</td>
<td>18</td>
</tr>
<tr>
<td>1.2.2 The talker-oriented perspective</td>
<td>21</td>
</tr>
<tr>
<td>1.2.2.1 Evidence of egocentric language use</td>
<td>22</td>
</tr>
<tr>
<td>1.2.2.2 Cascading activation</td>
<td>24</td>
</tr>
<tr>
<td>1.2.3 The passive evolutionary perspective</td>
<td>28</td>
</tr>
<tr>
<td>1.2.3.1 Evidence of language being optimized for communication</td>
<td>29</td>
</tr>
<tr>
<td>1.2.3.2 Exemplar dynamics</td>
<td>30</td>
</tr>
<tr>
<td>1.3 Research questions</td>
<td>32</td>
</tr>
</tbody>
</table>
1.4 Theory of mind assessment ........................................... 32
  1.4.1 False belief ...................................................... 33
  1.4.2 Other aspects of theory of mind ............................... 36
  1.4.3 Alternatives to false belief tasks ............................. 38
  1.4.4 The Autism-spectrum Quotient .............................. 43
1.5 Overview of thesis ................................................. 45

2 Theory of mind and phonetic reduction ............................ 46
  2.1 Introduction ..................................................... 46
  2.1.1 Predictability-based reduction ................................ 46
  2.1.2 Theoretical perspectives ..................................... 50
  2.1.3 Theory of mind ................................................. 55
  2.1.4 ToM assessment ............................................... 56
  2.1.5 Effects of autistic traits on linguistic variables .......... 60
  2.2 Experiment 1a: Semantic predictability-based reduction and second mention reduction ........................................... 63
    2.2.1 Introduction ............................................... 63
    2.2.2 Method ..................................................... 63
      2.2.2.1 Participants ........................................... 63
      2.2.2.2 Procedure ............................................ 63
      2.2.2.3 Acoustic measurements ............................... 66
      2.2.2.4 Analysis ............................................... 67
      2.2.2.5 Predictions ............................................ 68
    2.2.3 Results ..................................................... 69
      2.2.3.1 Individual difference scores ......................... 69
      2.2.3.2 Phoneme classification results ...................... 70
      2.2.3.3 Semantic predictability-based reduction results .... 72
      2.2.3.4 Second mention reduction results .................. 77
    2.2.4 Discussion .................................................. 80
      2.2.4.1 Equivalence of individual difference scores ....... 80
      2.2.4.2 Phonetic reduction and theory of mind ............. 81
      2.2.4.3 Consciousness and control .......................... 83
  2.3 Experiment 1b: Semantic predictability-based reduction and clear speech ......................................................... 85
    2.3.1 Introduction ............................................... 85
    2.3.2 Methods ..................................................... 85
      2.3.2.1 Participants ........................................... 85
      2.3.2.2 Procedure ............................................ 86
      2.3.2.3 Measurements ........................................ 86
      2.3.2.4 Analysis ............................................... 86
    2.3.3 Results ..................................................... 87
      2.3.3.1 Individual differences scores ....................... 87
      2.3.3.2 Word duration ....................................... 88
      2.3.3.3 Vowel duration ...................................... 90
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.3.4 Vowel dispersion</td>
<td>91</td>
</tr>
<tr>
<td>2.3.4 Discussion</td>
<td>92</td>
</tr>
<tr>
<td>2.4 Experiment 2: Word-specific VOT effects</td>
<td>94</td>
</tr>
<tr>
<td>2.4.1 Introduction</td>
<td>94</td>
</tr>
<tr>
<td>2.4.2 Methods</td>
<td>94</td>
</tr>
<tr>
<td>2.4.2.1 Participants</td>
<td>94</td>
</tr>
<tr>
<td>2.4.2.2 Materials</td>
<td>95</td>
</tr>
<tr>
<td>2.4.2.3 Procedure</td>
<td>95</td>
</tr>
<tr>
<td>2.4.2.4 Measurements</td>
<td>97</td>
</tr>
<tr>
<td>2.4.2.5 Analysis</td>
<td>97</td>
</tr>
<tr>
<td>2.4.3 Results</td>
<td>98</td>
</tr>
<tr>
<td>2.4.4 Discussion</td>
<td>99</td>
</tr>
<tr>
<td>2.5 Experiment 3: Lexical neighborhood density and frequency</td>
<td>102</td>
</tr>
<tr>
<td>2.5.1 Introduction</td>
<td>102</td>
</tr>
<tr>
<td>2.5.2 Methods</td>
<td>102</td>
</tr>
<tr>
<td>2.5.2.1 Materials</td>
<td>102</td>
</tr>
<tr>
<td>2.5.2.2 Procedure</td>
<td>102</td>
</tr>
<tr>
<td>2.5.2.3 Measurements</td>
<td>103</td>
</tr>
<tr>
<td>2.5.2.4 Participants</td>
<td>103</td>
</tr>
<tr>
<td>2.5.2.5 Analysis</td>
<td>103</td>
</tr>
<tr>
<td>2.5.3 Results</td>
<td>104</td>
</tr>
<tr>
<td>2.5.3.1 Individual difference scores</td>
<td>104</td>
</tr>
<tr>
<td>2.5.3.2 Word duration</td>
<td>104</td>
</tr>
<tr>
<td>2.5.3.3 Vowel duration</td>
<td>107</td>
</tr>
<tr>
<td>2.5.3.4 Vowel dispersion</td>
<td>108</td>
</tr>
<tr>
<td>2.5.4 Discussion</td>
<td>109</td>
</tr>
<tr>
<td>2.6 General Discussion</td>
<td>111</td>
</tr>
<tr>
<td>2.6.1 Summary of experiments</td>
<td>111</td>
</tr>
<tr>
<td>2.6.2 Phonetic reduction</td>
<td>113</td>
</tr>
<tr>
<td>2.6.3 Individual differences</td>
<td>114</td>
</tr>
<tr>
<td>2.6.4 Listener-oriented theories</td>
<td>115</td>
</tr>
<tr>
<td>2.7 Conclusion</td>
<td>116</td>
</tr>
</tbody>
</table>

3 Individual differences in the perception of phonetically reduced speech | 118 |
| 3.1 Introduction | 118 |
| 3.1.1 Effects of reduction in perception | 119 |
| 3.1.2 Individual differences in perception | 122 |
| 3.1.2.1 Speech perception in adverse conditions | 122 |
| 3.1.2.2 Speech perception under cognitive load | 124 |
| 3.1.2.3 Interim summary | 125 |
| 3.2 Experiment 4: Speech intelligibility in noise and clarity judgements | 127 |
| 3.2.1 Introduction | 127 |
| 3.2.2 Method | 128 |
List of Tables

2.1 Summary of reported effects of neighborhood density on vowel dispersion from three studies. A checkmark signifies a significant effect was observed, a cross signifies the lack of an effect. Empty cells indicate the study did not consider that column’s vowel. ................................. 50

2.2 Correlation matrix of the individual difference scores in experiment 1a, also including p-values for the significance of each pairwise comparison. .................. 70

2.3 Coefficients and significance values for logistic model predicting response in the phoneme classification task. ........................................ 71

2.4 Model output for word duration model in the sentence production task. .. 74

2.5 Model output for vowel duration model in the sentence production task. .. 76

2.6 Model output for vowel dispersion model in the sentence production task. . 77

2.7 Output of word duration model in the paragraph-reading task. .............. 78

2.8 Output vowel duration model in the paragraph-reading task. ............... 78

2.9 Output of vowel dispersion model in the paragraph-reading task. .......... 79

2.10 Correlation matrix of the individual difference scores in experiment 1b, also including p-values for the significance of each pairwise comparison. ...... 88

2.11 Output for the word duration model for experiment 1b. .................... 89

2.12 Output for the vowel duration model for experiment 1b. ................. 91

2.13 Output for the vowel dispersion model for experiment 1b. ............... 92

2.14 Example stimulus words presented to participants in each of the three conditions of experiment 2. ................................................. 96

2.15 Correlation matrix of the individual difference scores in experiment 2, also including p-values for the significance of each pairwise comparison. ...... 99

2.16 Output for the optimal model predicting VOT from the experiment 2 results. 99

2.17 Correlation matrix of the individual difference scores in experiment 4, also including p-values for the significance of each pairwise comparison. ...... 105

2.18 Output for word duration model, experiment 3. ............................ 105

2.19 Output for vowel duration model, experiment 3. ........................... 108

2.20 Output for vowel dispersion model, experiment 3. ........................ 109

2.21 Summary of the four reduction types examined in experiments 1a, 1b, and 3, and their interactions with individual differences in theory of mind. ...... 115

3.1 Number of stimuli in each cell of the $2 \times 2 \times 2$ design of the intelligibility and clarity experiment. ......................................................... 129
3.2 Model output for accuracy on intelligibility task. ........................... 133
3.3 Model output for clarity rating task. ........................................... 136
3.4 Model output for lexical decision accuracy. ............................... 149
3.5 Model output for lexical decision reaction times from word offset. .... 152
3.6 Model output for sentence acceptability response accuracy. ............ 165
3.7 Model output for sentence acceptability reaction times. ................. 168
3.8 Summary of effects of temporal reduction on the dependent variables in each of the perceptual tasks used in this chapter. ....................... 175

4.1 Correlation matrix of the three (log-transformed) word predictability measures in Clopper’s corpus. ........................................ 210
4.2 Output for the trigram models for duration difference and dispersion difference in the Clopper corpus. ................................. 215
4.3 Deletion rates (%) for each phoneme in the Buckeye corpus. ............ 222
4.4 Output of Poisson regression model predicting number of phoneme deletions per word in the Buckeye corpus. ......................... 222
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>20</td>
</tr>
<tr>
<td>1.2</td>
<td>34</td>
</tr>
<tr>
<td>1.3</td>
<td>38</td>
</tr>
<tr>
<td>1.4</td>
<td>41</td>
</tr>
<tr>
<td>1.5</td>
<td>42</td>
</tr>
<tr>
<td>2.1</td>
<td>52</td>
</tr>
<tr>
<td>2.2</td>
<td>59</td>
</tr>
<tr>
<td>2.3</td>
<td>65</td>
</tr>
<tr>
<td>2.4</td>
<td>72</td>
</tr>
<tr>
<td>2.5</td>
<td>73</td>
</tr>
<tr>
<td>2.6</td>
<td>75</td>
</tr>
<tr>
<td>2.7</td>
<td>76</td>
</tr>
<tr>
<td>2.8</td>
<td>76</td>
</tr>
<tr>
<td>2.9</td>
<td>90</td>
</tr>
<tr>
<td>2.10</td>
<td>101</td>
</tr>
</tbody>
</table>
2.11 Word duration as a function of number of lexical neighbors. Points indicate means of individual word types, and bars the standard error. ................................................. 106

2.12 Word duration as a function of talker RMITE score, split by word frequency and neighborhood density. See text for details on binning. ................................................. 107

3.1 Histogram of clarity judgement responses. ................................................................. 132

3.2 Word intelligibility as a function of duration, split by talker AQ scores. Curves are best fit binomials, grey bar is bootstrapped 95% confidence interval. .................................................. 134

3.3 Word intelligibility as a function of duration, split by talker word production condition. Curves are best fit binomials, grey bar is bootstrapped 95% confidence interval. .................................................. 135

3.4 Perceived word clarity as a function of duration, split by talker AQ score. Curves are best fit Poisson regression, grey bar is bootstrapped 95% confidence interval. .................................................. 137

3.5 Perceived word clarity as a function of duration, split by talker word production condition. Curves are best fit Poisson regression, grey bar is bootstrapped 95% confidence interval. .................................................. 138

3.6 Proportion correct response in the lexical decision task as a function of listener AQ score, split by stimulus production context. Note the outlier with AQ > 130 whose poor accuracy with HP targets shifts the overall slope of the trend. ........................................................................... 145

3.7 Predicted reaction time as a function of word duration under four different linear models. ................................................................................................................................. 147

3.8 Lexical decision accuracy on target words as a function of word duration. .................. 150

3.9 Lexical decision accuracy on target words as a function of word intelligibility. .......... 151

3.10 Lexical decision reaction times to correct responses to target words as a function of word intelligibility. .......................................................... 152

3.11 Lexical decision reaction times to correct responses to target words as a function of word duration, split by block. .................................................. 153

3.12 Lexical decision reaction times to correct responses to target words as a function of talker AQ score, split by block. .................................................. 154

3.13 Lexical decision reaction times to correct responses to target words as a function of talker AQ, split by word production context. Points show talker means by context, and standard errors. .................................................. 155

3.14 Lexical decision reaction time as a function of listener AQ score, split by word duration. Points indicate subject mean and standard error split by duration. .................................................. 156

3.15 Lexical decision reaction times to correct responses to target words as a function of word duration. Left panel shows simulated data with model parameters $b = -0.69, c = 733$; right panel shows linear trend lines for the observed data. .................................................. 157
3.16 Sentence acceptability accuracy as a function of talker AQ score, split by sentence context. Points indicate talker means and standard errors. 166

3.17 Boost in sentence acceptability accuracy between HP and LP words as a function of talker RMITE score, binned by word duration in milliseconds. See text for details. 167

3.18 Reaction times in the sentence acceptability task as a function of final word duration, split by sentence context. Points indicate item mean and standard error. 169

3.19 Reaction times in the sentence acceptability task as a function of listener AQ score. Points indicate subject mean and standard error. 170

4.1 Number of neighbors plotted as a function of log frequency with linear trend overlaid (left panel), and as a function of number of phonemes (right panel). 193

4.2 Contour plots of predicted vowel duration from $f_1$, the smoothed function of frequency and density (top panel), and $f_2$, the smoothed function of frequency and density conditioned on number of phonemes (bottom panel). 199

4.3 Contour plots of predicted vowel dispersion from the smoothed function of frequency and density. 200

4.4 Contour plots of predicted vowel duration from smoothed functions of frequency and density (top panel) and frequency and density conditioned on number of phonemes (bottom panel). 207

4.5 Dispersion difference in ERB between first and second word mentions as a function of the difference between log conditional trigram probability of the word pair. Points depict word means and standard errors. 216

4.6 Mean number of deletions per word as a function of lexical frequency, split by neighborhood density. The different panels show trends in words of differing numbers of phonemes. 223

4.7 Mean number of deletions per word as a function of number of mentions, split by neighborhood density. The different panels show trends in words of differing numbers of phonemes. 224

5.1 Vowel duration in the Clopper corpus as a function of log word frequency, split by speech style. Linear trend shown. 244
Chapter 1

Framing the problem

1.1 Predictability effects in speech production

A predictability effect can be said to exist for some language element if a relationship exists between the element’s predictability and the production of that element. By ‘element’ is meant any linguistic item, at various levels of analysis. Potential ‘elements’ include phones, syllables, words, and even syntactic phrases. Here, ‘predictability’ refers to the probability of occurrence given the context.

Predictability, and in particular the context it is defined over, can be parameterized in a number of ways. No context can be considered at all, in which case predictability is equivalent to unigram frequency—*the* is more common, and *cromulent* is less common. Thus, given no context, *the* is highly predictable, relative to other words, while *cromulent* is quite unpredictable, relative to other words. The simple context of preceding words or phonemes can be considered, which gives rise to \( n \)-gram models of predictability. Under such a model, *the* is predictable in some contexts (e.g. *and the*), but not in others (e.g. *this the*). Using semantics- and discourse-sensitive contexts can lead to other models of predictability—for instance, in this paragraph, the word *predictability* is more predictable than it is in a randomly-sampled paragraph from a book on literary theory. Finally, rather than predicting a word based on knowledge of its context, we can try to predict a word based on partial knowledge of the word’s
structure—for example, if we know the word is a monosyllable ending in /ræft/, then predictable words are raft (the degenerate case), graft, craft, and draft. None of these models are true and complete reflections of human predictability modeling, but each of them can serve as useful approximations of reality.

The definition of ‘predictability effect’ given above made reference to the production of linguistic elements. In this dissertation, production is mainly measured through the acoustic signal. Predictability effects, by definition, influence some dimension of production, and thus some acoustic dimension(s). Examples include changes to word duration, vowel duration, vowel dispersion in the formant space, and other timing- or spectrum-related changes. When these changes reduce the acoustic distinctiveness of an element, by minimizing phonetic substance (e.g. temporal shortening) or prominence (e.g. vowel centralization), the change is referred to as one of acoustic reduction. Throughout this dissertation, I assume a link between predictability effects and acoustic reduction.

1.1.1 Observed predictability effects and their properties

As discussed above, multiple contexts for parameterizing predictability are possible. This section addresses how the production of linguistic units can vary in relation to their predictability—i.e. the nature of predictability effects. The general trend observed throughout is that predictable elements are produced in a reduced manner. As such, we expect that highly predictable words will be pronounced with shorter word durations, shorter vowel durations, and less disperse vowels than less predictable words.
1.1.1.1 Lexical frequency

Lexical frequency is one of the oldest predictability effects documented, dating to at least Zipf's (1929) observation that word length is negatively correlated with frequency. Frequency is usually defined as the number of times a word type appears in some corpus. Lexical frequency has long been acknowledged to be a relevant factor in psycholinguistics, with high frequency words tending to be named faster than low-frequency words (Damian, 2003; Griffin and Bock, 1998; Jescheniak and Levelt, 1994), and in other domains such as language acquisition (see Ellis, 2002, for review).

In terms of acoustic reduction, high-frequency words have been observed to have shorter word and vowel durations (Arnon and Cohen Priva, 2013; Aylett and Turk, 2004; Bell et al., 2009; Fidelholtz, 1975; Gahl, 2008; Gahl et al., 2012; Goldrick and Blumstein, 2006; Kapatsinski, 2010; Munson and Solomon, 2004; Myers and Li, 2009; Pate and Goldwater, 2011; Pluymaekers et al., 2005b; Tomaschek et al., 2013) and less disperse vowels (Munson, 2007; Munson and Solomon, 2004) than low-frequency words.

The results from articulatory studies are less uniform. Lin et al. (2014) used ultrasound to investigate the anterior constriction during the lateral portion of English /VC/ sequences, where the last C was a labial or velar. The lateral phoneme is often velarized in English, with minimal coronal (anterior) constriction. They found that high-frequency words like help had much smaller degrees of constriction than low-frequency words like whelp—in other words, the articulation of the high-frequency words was reduced relative to the low-frequency words. By contrast, Tomaschek et al. (2013) used electromagnetic articulography to investigate the vowels /a, a:, i, i:/ in German, and found that vowels in high-frequency words tend to be produced with greater tongue movement than those in low-frequency words—that is, the low-frequency words were
reduced relative to the high-frequency words. Additionally, their acoustic data suggested that while high frequency leads to temporal reduction for the phonologically short vowels, it leads to temporal enhancement for the phonologically long vowels. That is, the short vowels get shorter, while the long vowels get longer. Tomaschek et al. (2013) interpreted their results in terms of frequency of use leading to more well-trained and precise articulatory movements. Subsequent analysis by Tomaschek et al. (2014), however, revealed a substantial influence of segmental context around the vowel, and they concluded that the data do not suffice to make firm conclusions about the relationship between frequency and articulation. The effects of frequency on articulation are clearly not well understood; for this reason, the present study focuses purely on the acoustic consequences of reduction.

1.1.1.2 Phonological neighborhood density

Phonological neighborhood density is a measure of how many words are similar to a given word in a language. For example, the word *cat* has many phonological neighbors, such as *hat*, *mat*, *cot*, and *cap*, and is therefore considered to be in a dense neighborhood. The word *clasp*, on the other hand, has relatively few neighbors, and is therefore considered to be in a sparse neighborhood.

Early inklings of the importance of this concept for speech perception were discussed by Savin (1963), who noted that if listeners are not able to make out the identity of a word, they may still be able to perceive some of the word's properties, such as the number of syllables it has or what the vowels are. Therefore, if a word is very distinctive given its phonological makeup, and thus has few lexical competitors, it is more likely to be accurately identified than a less distinctive word with many competitors. The number of possible competitors—the word's neighborhood—directly constrains the listeners' possible accuracy in their identification of the word. In line
with this prediction, Savin observed higher intelligibility rates for words with greater numbers of syllables, which tend to have fewer neighbors. However, a formal definition of neighborhood or similarity was not provided.

Landauer and Streeter’s (1973) study marked one of the first attempts to directly quantify phonological neighborhood density. They conceived of similarity as the amount of information loss required to render two words indistinguishable. For the purposes of their study, they defined a word’s neighbors in terms of orthography. A word’s neighbors are all the words that can be generated by substituting a single letter in that word. Therefore, the word base has the neighbors case, ease, vase, bade, and bass, among others. Although somewhat crude and limited to comparing words of the same orthographic length, this simple metric enabled Landauer and Streeter to establish basic distributional properties of the English lexicon.

Building on Landauer and Streeter’s work, Pisoni et al. (1985) and Luce (1986) developed the now-standard method of computing neighborhood density: given a string of phonemes, the neighbors are all the words whose phoneme strings are one substitution, addition, or deletion away from the original string. With this metric, base has the neighbors case, brace, ace, among others. Note that ease is no longer a neighbor under this definition. Luce (p6) noted that this definition makes “strong assumptions” about similarity, viz. that all phonemes are equally similar and that all positions within a word are equally important. These assumptions are clearly false (Miller and Nicely, 1955), but they simplify the process of calculating the metric. This definition’s main advantages are that it is easy to compute, requiring only a phonemically-transcribed dictionary, and gives neighborhood estimates which accord with both intuition and the experimental evidence.

This metric was used to test several predictions of models of word recognition, such as the Neighborhood Activation Model (Luce and Pisoni, 1998) and Cohort
theory (Marslen-Wilson and Tyler, 1980; Marslen-Wilson and Welsh, 1978). As predicted, neighborhood density was observed to be influential in a number of perceptual domains: Luce and Pisoni (1998) found that performance in an auditory lexical decision task is worse for words in high-density neighborhoods compared to words in low-density neighborhoods. Several subsequent studies support the generalization that higher-density words are harder for the perceptual system to process than lower-density words (e.g. Bradlow and Pisoni, 1999; Garlock et al., 2001; Luce et al., 2000; Sommers and Danielson, 1999; Vitevitch and Luce, 1999; Vitevitch et al., 2008).

Additionally, effects of neighborhood density on word production have been noted. Curiously, the generalization that arises from these studies is that, relative to low-density words, high-density words are easier to access for production, the opposite generalization to that of the perception studies. For example, relative to words in sparse neighborhoods, words in denser neighborhoods tend to have a shorter naming latency—that is, the length of time from visual display to the onset of speech in a naming task (Dell and Gordon, 2003; Heller and Goldrick, 2014; Vitevitch, 2002). Similarly, Gordon (2002) and Stemberger (2004) have presented evidence that high-density words are produced with fewer speech errors than low-density words. These findings are consistent with an account whereby high-density words have higher activation levels than low-density words, due to competition from their neighbors (e.g. Baese-Berk and Goldrick, 2009). This activation leads to the faster naming latencies and fewer production errors; such an account also predicts that high-density words ought to be hyperarticulated relative to low-density words. This prediction is borne out: words in denser neighborhoods have been observed to be produced with more expanded vowel spaces than words in sparser neighborhoods (Clopper and Tamati, 2014; Munson, 2007; Munson and Solomon, 2004; Scarborough, 2004, 2010; Watson and Munson, 2007, 2008; Wright, 2004). Additionally, there exists some evidence that
words in dense neighborhoods have longer word and vowel durations (Burdin and Clopper, 2015; Burdin et al., 2014a,b; Kryuchkova and Tucker, 2012; Scarborough, 2010) than words in sparse neighborhoods, but this finding has not been consistently observed (Munson and Solomon, 2004).

In addition to effects on vowels and word duration, neighborhood density has been argued to influence consonant production. The classic study in this regard is that of Baese-Berk and Goldrick (2009), who demonstrated that words like cod, which have a minimal pair neighbor god, are produced with a longer initial VOT than words like cop, which do not have a minimal pair neighbor *gop. (This study is discussed in more detail in Section 1.2.2.) Although Baese-Berk and Goldrick (2009) only considered minimal pair status, subsequent work has interpreted this distinction in terms of neighborhood density, and the result is consistent with the research on neighborhood effects on vowels: words in denser neighborhoods are hyperarticulated in the relevant dimension of contrast (cop has a longer VOT), and words in sparser neighborhoods are hypoarticulated (cod has a shorter VOT). These results have been replicated several times (Bullock-Rest et al., 2013; Fox et al., 2015; Goldrick et al., 2013; Kirov and Wilson, 2012; Peramunage et al., 2011).

In the perception literature, then, it has been established that, relative to words in sparse neighborhoods, words in dense neighborhoods are difficult to access, while in the production literature, words in dense neighborhoods are easier to access. This inconsistency between production and perception has led some researchers (Fox et al.,
2015; Gahl, 2015; Gahl et al., 2012; Yao, 2011) to suggest that the observed production effects are largely due to uncontrolled contextual factors which happen to correlate with density. For instance, it is well-established that number of neighbors correlates positively with both word frequency and segment bigram probability (Frauenfelder et al., 1993), and this collinearity could interfere with the interpretation of purported density effects in production if not adequately controlled. Gahl et al. (2012) and Gahl (2015) have further suggested that once these factors are controlled for, words in denser neighborhoods in fact have less expanded vowel spaces than words in sparser neighborhoods, consistent with the notion from the perception literature that words in dense neighborhoods are in fact more difficult for the speaker to access.

Although the majority of work on neighborhood density thus far has focused on English, several of the major findings have been replicated in other languages. These languages include Cantonese (Kirby and Yu, 2007), Basque (Arbesman et al., 2010), Dutch (Frauenfelder et al., 1993), Hawaiian (Arbesman et al., 2010), Korean (Holiday and Turnbull, 2015), Mandarin (Arbesman et al., 2010; Tsai, 2007; Yip, 2002), Norwegian (Johnsen, 2011), and Spanish (Arbesman et al., 2010; Baus et al., 2008; Vitevitch and Rodríguez, 2004; Vitevitch and Stamer, 2006, 2009). Additionally, work on neighborhood density has been extended to accented and L2 speech (e.g. Imai et al., 2005).

1.1.1.3 Discourse mention

First discussed in detail by Chafe (1974), discourse mention refers to whether a word has been mentioned previously in the discourse. Generally speaking, words that have been mentioned already in the discourse are presumed to be more accessible and more predictable than words that have not been mentioned. This reasoning follows from the idea that words that have been mentioned are established in the common ground.
Acoustic work has established that the second mention of a word generally has a shorter word duration, a shorter vowel duration, and is less intelligible in isolation than the word's first mention (Baker and Bradlow, 2009; Bard and Anderson, 1994; Bard et al., 2000, 1989; Burdin and Clopper, 2015; Fowler, 1988; Fowler and Housum, 1987; Fowler et al., 1997; Galati and Brennan, 2010; Hawkins and Warren, 1994; Kahn and Arnold, 2012, 2015; Kaiser et al., 2011; Lam and Watson, 2010, 2014; Pate and Goldwater, 2011; Sasisekaran and Munson, 2012; Shields and Balota, 1991; Turnbull, 2015; Vajrabhaya and Kapatsinski, 2011). It has also been suggested that second mentions are less likely to bear a (prominent) pitch accent than first mentions (Baker and Bradlow, 2009; Burdin and Clopper, 2015). There is also evidence that second mention reduction is linked to other aspects of communication: Hoetjes et al. (2015) observed that co-speech gesturing which accompanies second mentions tends to be reduced in magnitude relative to gesturing which accompanies first mentions. Similarly, Hoetjes et al. (2012) documented second mention reduction effects in Dutch Sign Language.

1.1.1.4 Semantic predictability

Semantic predictability is a broad term and refers to various non-lexical factors which influence the predictability of a word in a given context. Lieberman (1963) found that words which are predictable given the preceding context were less intelligible in isolation than unpredictable words. Acoustically, this distinction has been observed to manifest itself in shorter word and vowel durations (Aylett and Turk, 2006; Bell et al., 2009; Clopper and Pierrehumbert, 2008; Engelhardt and Ferreira, 2014; Ernestus et al., 2015; Gahl and Garnsey, 2004; Hunnicutt, 1985, 1987; Jurafsky et al., 2001; Lieberman, 1963; Moore-Cantwell, 2013; Pate and Goldwater, 2011; Pluymaekers et al., 2005a; Tily and Kuperman, 2012), less disperse vowels (Aylett and Turk, 2006; Clopper and Pierrehumbert, 2008; Jurafsky et al., 2001), and less prosodic prominence (Kaland
et al., 2014; Turnbull, 2015; Watson et al., 2008) in predictable words relative to unpredictable words.

1.1.1.5 Other predictability effects

There are a wide range of other phenomena related to predictability which are not considered in this dissertation. Some phenomena relate to alternative definitions of predictability. For example, Cohen Priva’s (2008) measure of ‘informativity’ quantifies how much information, on average, a word contributes to a sentence. Seyfarth (2014) investigated the relationship between informativity and reduction and found that more informative words have shorter durations than less informative words, even after controlling for word frequency. Van Son and Pols (2003) considered the ‘information content’ of phonemes—a measure of how unlikely the phoneme is, given its phonological and lexical context—and found that phonemes with a higher information content (i.e. phonemes that are less likely) were produced with longer duration and a higher spectral center of gravity than low information content phonemes. This result is similar in substance to the finding that phonemes from unstressed syllables in Dutch have a lower center of gravity than phonemes from stressed syllables (Van Son and Pols, 1999). To the extent that a lower spectral center of gravity reflects decreased articulatory effort (de Jong, 1995; Sluijter and van Heuven, 1996), these results accord with the idea that predictable elements are produced in a reduced manner.

Predictability effects have also been demonstrated in the phonological domain, where categorical processes such as voicing or deletion are considered. Cohen Priva’s (2008) informativity measure, mentioned above, negatively correlates with likelihood of phoneme deletion—that is, phonemes which are less informative are more likely to be deleted. Similarly, English coronal stop deletion is more common in words with a
higher conditional probability than a lower one (Coetzee and Kawahara, 2013; Raymond et al., 2006). Voicing contrasts in Catalan are more likely to be completely neutralized when the word is in a predictable context than in an unpredictable context (Charles-Luce, 1993), and devoicing of geminates in Japanese loanwords happens more often to higher frequency words (Coetzee and Kawahara, 2013). Conditional probability of words in context predicts variable schwa epenthesis in Dutch; words in a less predictable position are more likely to undergo epenthesis than more predictable words (Tily and Kuperman, 2012). Information density predicts presence or absence of relativizer that in relative clauses, such that a high density (low predictability) sequence is more likely to have that, i.e. more phonological material, than a low density (high predictability) sequence (Jaeger, 2010; Levy and Jaeger, 2007). See also Frisch (2011) for a review of frequency effects in phonotactic judgements. As this list of observed effects shows, the phonological phenomena (e.g. devoicing, deletion), the predictability measures (e.g. informativity, frequency), and the level of analysis (e.g. phoneme, word) are all quite diverse, yet there is remarkable consistency in the generalization that more predictable elements have less phonetic substance or prominence.

1.1.2 The source of predictability effects

As reviewed above, predictability effects are diverse phenomena, and there are many ways to define predictability. The next section summarizes three major theoretical perspectives of these effects. Each of these theories assumes that predictability effects can be reduced to a single relevant factor. The theories differ in what the factor actually is, but each theory holds, either implicitly or explicitly, that the single factor is sufficient to account for predictability. This perspective is questioned in Chapter 4 of this dissertation, where analysis of phonetic corpora is utilized to trace the source(s) of
predictability-based phonetic reduction. To foreshadow the results, the modeling sug-
tests that reduction is best considered as arising from an interaction between at least
two and possibly three factors: lexical factors, such as lexical frequency and neigh-
borhood density; and contextual factors, such as semantic predictability and discourse
mention. A further division between discourse-specific factors (such as discourse men-
tion) and domain-general factors (such as semantic predictability) is possible. This
result is consistent with ‘hybrid’ models of reduction which allow multiple sources of
reduction, such as that put forth by Watson (2010), discussed in detail in Chapter 4.

1.2 Theoretical perspectives

Predictability effects in speech production have been explained from three primary
theoretical perspectives: listener-oriented, talker-oriented, and passive evolutionary
constraints. This section summarizes these three approaches, their primary similarities
and differences, and the empirical evidence used to support them. From each perspec-
tive, one specific model that has been proposed in the literature is described in detail,
and the kinds of data that would be necessary to falsify each model are discussed.

1.2.1 The listener-oriented perspective

According to the listener-oriented perspective (Aylett, 2000; Aylett and Turk, 2004,
2006; Frank and Jaeger, 2008; Galati and Brennan, 2010; Genzel and Charniak, 2003;
Jaeger, 2013; Jaeger and Tily, 2011; Pate and Goldwater, 2015; Qian and Jaeger, 2012;
Ramscar and Baayen, 2013; Schober, 1993; Turk, 2010; Van Son and Pols, 2003; Van
Son and Van Santen, 2005, inter alia), predictability effects serve a functional pur-
pose: enhancing communicative success while minimizing talker effort. Some sounds
or words are more likely to be misperceived by the listener than other sounds or words.
This misperception could be due to acoustic-perceptual factors (such as masking of acoustic cues in certain phonological contexts), or to linguistic predictability factors (such as the likelihood of an adjective following a noun). Under the listener-oriented perspective, talkers are aware of these potential comprehension difficulties, and accordingly make efforts to enhance the acoustic prominence of these ‘difficult’ sounds or words. On the other hand, the talker is free to phonetically reduce ‘easy’ sounds or words, which are likely to be perceived correctly by the listener. By hypothesis, it is easier for the talker to produce such reduced variants than to produce the enhanced variants, which is why the reduced variants are common when the listener’s successful perception is likely. The listener-oriented perspective, then, claims that predictability effects exist to facilitate successful perception by the listener, and to ease the articulatory burden on the talker.

Two assumptions are implicit in the preceding discussion: first, that reduced speech is easier for the talker to produce than unreduced speech; second, that reduced speech is harder for the listener to perceive than unreduced speech. If either of these assumptions fail, so too does the listener-oriented account. If reduced speech is not actually easier to produce than unreduced speech, then talkers have no incentive to reduce; if reduced speech is not actually harder to perceive than unreduced speech, then talkers have no incentive to fail to reduce. In terms of production, the notion of ‘articulatory effort’ has a long but blurry history within phonetics. There is no currently-accepted standard of measurement, and several researchers have suggested that there are actually multiple dimensions of effort (Boersma, 1998; Cheng and Xu, 2015; Pouplier, 2003). Nevertheless, the idea that acoustically reduced productions—which have shorter durations and less peripheral vowel productions—require less articulatory effort is intuitively reasonable. This assumption is adopted in this dissertation. The
second assumption, however, is the topic of Chapter 3, where the assumption of a perceptual benefit for unreduced speech is examined in detail.

1.2.1.1 Evidence for listener-orientation

Much of the evidence cited in section 1.1.1 on the nature of predictability effects is consistent with all three perspectives. In this section, and those following, I will focus on research which directly bears on the question of teasing these perspectives apart, such that it exists.

In addition to broad intuitive appeal, a number of studies have provided empirical evidence which is consistent with the listener-oriented perspective. For example, in a corpus study of second mention reduction, Bard et al. (1989) found that many of the second mentions analyzed were not actually reduced relative to their first mention; indeed, some were longer. A closer look revealed that many of these ‘second’ mentions, although they were the second time the wordform was uttered, were in fact referring to a different referent in the real world. In terms of discourse, the information being provided by these words was actually new, not given, as had been assumed in the preliminary analysis. A reanalysis, considering whether the information conveyed was given or new, revealed that given items were reduced relative to the new items. Rather than simply keeping track of which words had already been said, talkers must have had a discourse model updated in real-time, regarding which referents were salient in the conversation.

Another kind of listener modeling is variation in speech style, which can vary according to communicative scenario. The style of speech adopted when speaking with someone with a hearing impairment can be quite different from that used when addressing a friend with normal hearing (Smiljanić and Bradlow, 2009). The term ‘clear
speech’ is usually used in opposition to ordinary ‘plain speech’, to include speech directed to non-native speakers (Uther et al., 2007), people with hearing impairments (Smiljanić and Bradlow, 2009), infants (Cristià, 2010; Kuhl et al., 1997), and pets (Burnham et al., 2002; Xu et al., 2013). Characteristics of clear speech include temporal enhancement and increase in fundamental frequency range, and the speech is generally more intelligible than plain speech (Baker and Bradlow, 2009; Cho et al., 2011; Granlund et al., 2012; Kang and Guion, 2008; Picheny et al., 1985, 1986; Smiljanić and Bradlow, 2005, 2008).

Burnham et al. (2002) found that although infant-directed speech and pet-directed speech are superficially similar, they differ in important ways. Principally, the vowels of infant-directed speech are more disperse throughout the vowel space, while those of pet-directed speech are not as acoustically distinct. This result was interpreted as evidence that talkers are sensitive to their audience, and that talkers enhance their vowels when talking to children in order to help them learn the language. The pets, in this case cats and dogs, are not going to learn the language and thus the talkers do not need to help them do so. In a follow-up study, Xu et al. (2013) found that parrot-directed speech lies between dog-directed speech and infant-directed speech on a continuum of vowel hyperarticulation, confirming the notion that (perceived) linguistic capacity of the interlocutor influences speech style. However, this result could be explained in terms of attempting to get the interlocutor to imitate one’s speech, rather than in terms of communication or acquisition per se.

McMurray et al. (2013) examined infant-directed speech with a view to determining if the intelligibility benefit was an intrinsic feature of the speech style—which would suggest a didactic or communicative purpose to the speech—or if it was an epiphenomenon of the slower speech rate. Their analyses of vowel dispersion and voice onset time (VOT) showed that the enhanced formants and longer VOT values
observed in infant-directed speech could be attributed to speech rate alone, and they concluded that infant-directed speech does not appear to be crafted to specifically support learning. However, research into speech directed to hearing-impaired listeners has suggested that the slower rate of clear speech is somewhat orthogonal to the intelligibility benefit (Amano-Kusumoto et al., 2014; Krause and Braida, 2002, 2004). These studies have used methods specifically designed to elicit particular speech styles at particular rates. Other research into speech rate suggests that there are language-specific strategies for how segments are affected by rate. For example, while English VOTs decrease as a linear function of speech rate (slower speech leading to longer VOTs), VOTs in Catalan are consistent across speech rates (Solé, 2007). These findings suggest that talkers have fine control over these phonetic details. Taken together, the literature on speech rate suggests that McMurray et al.’s (2013) conclusions may be premature. Nevertheless, there are results from the infant-directed speech literature suggesting that a purely communicative or didactic account does not explain the entire phenomenon. Martin et al. (2014) examined infant-directed speech in Japanese, a language where vowels can devoice in particular environments. Their results demonstrated a lower rate of high vowel devoicing in infant-directed speech than plain speech, but paradoxically, a higher rate of mid vowel devoicing in the same context. Martin et al. (2014) interpreted these results as indicative of a complex interaction among competing factors, and suggested that the unidimensional continuum of hypo- and hyper-speech is inadequate to fully characterize speech style.

Clear speech effects and predictability effects are clearly distinct—clear speech is the result of an explicitly listener-oriented and conscious effort (Bradlow, 2002; Lindblom et al., 1992), while predictability effects appear to be the result of unconscious processes. The evidence suggests that clear speech effects and predictability effects are independent—in clear speech, less predictable words are still hyperarticulated relative
to more predictable words (Baker and Bradlow, 2009; Turnbull and Clopper, 2013). More research is needed to determine the relationship, if any, between these phenomena.

Although not directly investigating predictability effects, there are several additional lines of research which support the idea that language users strive for effective and efficient communication. Some English words have both a long form, like mathematics, and a short form, like math. Mahowald et al. (2013) presented English speakers with sentences like those in (1) and (2), and asked them to choose between math or mathematics to complete the sentence. Note that the context in (1) is predictive of the continuation, while the context of (2) is neutral or unpredictable.

(1) Susan was very bad at algebra, so she hated ...

(2) Susan introduced herself to me as someone who loved ...

The results demonstrated that the participants chose the short form in the predictable context (67%) more often than in the unpredictable context (56%). This finding was interpreted as evidence in favor of listener-oriented language use, whereby the talkers enhance (or fail to reduce) in an unpredictable context to aid intelligibility for the listener.

There is also evidence that talkers apply these listener-oriented strategies to novel communication paradigms. Fedzechkina et al. (2012) had participants learn an artificial language with fixed word order and optional case marking to indicate which noun was the subject and which was the object of a simple sentence. After learning the language, the participants produced novel utterances, based on videos they saw on-screen. The participants used case-marking more often when the utterance involved an atypical inanimate subject, versus a typical animate subject. In other words, the participants were using case-marking to highlight which noun was the subject and which
noun was the object in less typical—less predictable—situations. Since the language exhibited a fixed word order, such highlighting was redundant. This result was interpreted in terms of efficient communication, such that the participants were striving for both effective communication and conservation of effort.

Other studies involving communicative tasks have reported talkers making adjustments to their speech in response to listener characteristics. Rosa et al. (2015) observed talkers to produce longer words when addressing a distracted listener than an attentive listener, and Pate and Goldwater (2015) noted that speech is more intelligible when the interlocutors cannot see each other than when face-to-face (see also Bruce et al., 2013). These findings are consistent with the idea that talkers are monitoring their listeners in real-time and making adjustments as necessary.

Taken together, these results suggest that talkers are sensitive to the needs of their interlocutor, both real and imagined. The reviewed studies demonstrate a variety of adaptation mechanisms whereby talkers ‘help’ their interlocutor, generally by producing forms with more phonetic, phonological, or morphological content. This extra content makes utterances more redundant, which in turn makes the signal more robust to noise, thereby boosting the likelihood of the listener’s correctly receiving the message. Formulation of speech communication as a system of message sending, with signal redundancy as a central component, is a view adopted by the smooth signal redundancy hypothesis (Aylett, 2000; Aylett and Turk, 2004, 2006; Turk, 2010), to which we now turn.

1.2.1.2 Smooth signal redundancy

In Aylett and Turk’s (2004) smooth signal redundancy proposal, two competing constraints are argued to exist: those of reliable communication and conservation of effort. In this regard, the proposal is very similar in spirit to Lindblom’s (1990) hyp-
hyper-articulation (H&H) theory, which similarly situated talker behavior as a balance between hypo-articulation (which conserves effort) and hyper-articulation (which enhances communication). For communication to be reliable and robust, the signal needs to be clear enough for the message to be transmitted successfully. On the other hand, conservation of effort demands that we be lazy and do the bare minimum to get by. Too much of a focus on reliability and our speech is redundant; too much focus on brevity and we are not understood. ¹ This balance between redundancy and robustness is reminiscent of information theory, where the concerns about successful transmission of the message are similar. As Pierce (1981, p164) wrote in his introductory textbook on information theory:

The whole problem of efficient and error-free communication turns out to be that of removing from messages the somewhat inefficient redundancy which they have and then adding redundancy of the right sort in order to allow corrections of errors made in transmission.

According to Aylett and Turk (2004), redundancy in speech communication is of two kinds. One kind is language redundancy, which is broadly equivalent to the definitions of ‘predictability’ that have been discussed here. Parts of a message are more redundant, given how predictable those parts are. The other kind of redundancy is acoustic redundancy, which is conceptualized as the likelihood that the signal will be perceive correctly based on the acoustic properties alone. The sum of these two redundancies is the total signal redundancy. Aylett and Turk (2004) maintained that language users strive to ensure that the signal redundancy is smooth throughout an utterance; see Figure 1.1. This balance between redundancies accounts for the observed relationships between frequency and phonetic content. High frequency words

¹This problem was apparently very familiar to the Roman poet Horace, who wrote in Ars Poetica (25–26) that Brevis esse laboro, obscurus fio (‘I labor to be brief, and I become obscure’).
are high in linguistic redundancy, while low frequency words are low in linguistic redundancy. In order to achieve a signal redundancy which is smooth (i.e. constant), the high frequency words are accompanied by a low acoustic redundancy (i.e. acoustic reduction), while the low frequency words have a relatively high acoustic redundancy (i.e. no reduction, or enhancement).

Implicit in this model is the talker’s tacit awareness of what the listener considers predictable or unpredictable. For this awareness to be possible, the talker requires a theory of mind. Theory of mind is discussed in more detail in Section 1.4, but for now it will suffice to define it as the ability to recognize and model the mental states

![Figure 1.1: Schematization of the relationship between linguistic and acoustic redundancy. Acoustic redundancy is a function of the inverse of linguistic redundancy, resulting overall in a smooth signal redundancy, or a uniform information density. Adapted from Turk (2010).](image)
of others. By this reasoning, talkers with an impoverished theory of mind, who are less able to track their interlocutor’s mental state, should exhibit smaller or more inconsistent predictability effects than talkers with a fully-developed theory of mind, who are excellent at tracking their interlocutor’s mental state. If it is not the case that theory of mind is linked to predictability effects, then the evidence for this model is substantially undermined.

This model also predicts that reduction happens as a response to context. This prediction could be shown to be false if there existed evidence of a reduction that failed to apply in some predictable context, or of a reduction that over-applied in some unpredictable context. Finally, this model could be falsified by falsifying the assumption mentioned above: that articulating reduced forms is physiologically easier than articulating enhanced forms. If this were not the case, the logical basis for this model (that talkers reduce when they can get away with it) collapses.

Specifically, one of the questions this dissertation examines is the prediction relating to theory of mind. Chapter 2 addresses whether, as predicted, individual variation in theory of mind ability correlates with extent of phonetic reduction.

1.2.2 The talker-oriented perspective

According to the talker-oriented or ‘egocentric’ perspective, predictability effects are essentially epiphenomenal, arising from interactions in the cognitive architecture of the speech production system. The precise formulation and reasoning behind the effects is theory-specific, but the general theme is that ‘easy’ sounds or words are accessed or processed more quickly and more easily, which leads to a faster and less precise production. ‘Difficult’ words and sounds are accessed or processed less quickly and less easily, resulting in a more effortful and precise production. From this perspective, there
is no direct role of perceptibility or reference to the ability of the listener to understand the message.

1.2.2.1 Evidence of egocentric language use

There are several studies which suggest that language users can fail to take into account the perspective of their interlocutor, and instead rely on their own perspective to determine when and how predictability effects should be applied. Bard et al. (2000) carried out an investigation of second mention reduction, analyzing data from a map task corpus. After the instruction giver had finished guiding their partner through a map, their partner changed and they had to guide the new partner through the same map. All of the mentions of the landmarks were, in this context, discourse-given from the perspective of the instruction giver, but discourse-new from the perspective of the partner being led. Bard et al. (2000) found that the productions in this second trial were both shorter in duration and less intelligible in isolation than the productions from the first trial—suggesting that second mention reduction had taken place. Despite the instruction giver being aware that their interlocutor had changed and was not aware of the discourse context, the instruction giver still reduced the discourse-given tokens. Bard et al. (2000) interpreted this finding as evidence of an egocentric pattern, where the talker’s situational knowledge assumes primacy over their modeling of the listener’s knowledge.

Evidence from a different domain comes from a series of experiments carried out by Baese-Berk and Goldrick (2009). In these experiments, the experimental participant had to instruct a partner to click on an item on a computer display. Both the instructor and the partner saw the same display of items. In a condition where two of the displayed items were referents of a VOT minimal pair—e.g. cod and god—more extreme VOT values were observed on the target word than when the displayed items
were not minimal pairs. This enhancement of contrast in a potentially ambiguous context is consistent with a listener-oriented perspective. However, when the same target item \( cod \) was displayed without any minimal pair competitor, the VOT enhancement was still observed, albeit to a smaller degree. This enhancement of a perceptual cue in an unambiguous context cannot be accounted for by a listener-oriented perspective. Baese-Berk and Goldrick (2009) explained the findings in terms of activation levels and contextual priming, explained in more detail in Section 1.2.2.2.

Kapatsinski (2010) carried out a study of speech error repairs in the Switchboard corpus. In speech error repair, the incorrect word can either be interrupted mid-production, as in (3); or the repair can occur after the word has been fully produced, as in (4) (examples from Kapatsinski, 2010, p74). Analysis revealed that if the repaired word was high frequency, it was more likely to be fully produced, whereas if the repaired word was low frequency, it was more likely to be interrupted. This finding suggests that frequent words are harder to interrupt or otherwise inhibit, and therefore the production of frequent words is more automatic than that of infrequent words. This automaticity reflects the action of talker-internal mechanisms, and cannot be of benefit to the comprehension of the listener.

(3) It was \textit{pathe–}, I mean, it was \textbf{horrible}.

(4) That’s why we were surprised to see ‘\textit{Toyota}’ \textit{written}, I mean, \textbf{imprinted} on the engine.

One set of results which at first appears to support the listener-oriented rather than the talker-oriented perspective comes from the psycholinguistic literature on the prosodic marking of syntactically ambiguous sentences—e.g. PP attachment height. Early work by Allbritton et al. (1996) and Snedeker and Trueswell (2003) suggested that talkers use prosody to disambiguate only when both potential interpretations are possible,
given the context. When no confusion is possible, no disambiguating prosody is used. These findings supported Fox Tree and Meijer’s (2000) assertion that listeners do not make extensive use of such prosodic cues in comprehension. However, more recently, substantial counter-evidence has emerged. Kraljic and Brennan (2005), Schafer et al. (2000), and Speer et al. (2011) have provided evidence that talkers use prosodic disambiguation of syntactic phrasing regardless of the context and regardless of the needs of the listener—provided, of course, that the talker is actually aware that a potential ambiguity exists. These more recent experiments involved more naturalistic tasks and the speech was more spontaneous than those of Snedeker and Trueswell (2003), suggesting a higher level of ecological validity. Taken together, the evidence from syntactic disambiguation appears to in fact support a talker-oriented account of dialogue.

In summary, there is a variety of evidence suggesting egocentric patterns in language use. On the one hand, results like those reported by Bard et al. (2000) suggest that talkers can fail to take into account the perspective of their interlocutor and reduce in inappropriate ways. On the other hand, Baese-Berk and Goldrick’s (2009) results suggest that talkers reduce (and fail to reduce) even when there is no communicative imperative or potential for misperception. Both of these sets of results are consistent with a talker-oriented model whereby predictable elements are easier to access during speech planning (Kapatsinski, 2010), facilitating their phonetic reduction.

### 1.2.2.2 Cascading activation

The model outlined by Baese-Berk and Goldrick (2009) is a cascading activation model (Goldrick et al., 2011; Goldrick and Blumstein, 2006). This model is modular and feed-forward, such that processes happen sequentially and a later module cannot affect the previous modules—e.g. articulation cannot directly influence the lexicon in speech production. The cascading component of the model refers to the possibility of
multiple representations being simultaneously activated and fed forward to the next process. Goldrick and Blumstein (2006) used the example of production of the word *calf*. Lexical selection of this word entails partial activation of the word’s semantic neighbors, such as *cow* and *foal*. Activation of each of these lexical items is passed to the phonological level, where their corresponding phonological representations become activated. If all goes well, *calf* is the most highly activated representation and it is the word that is actually produced; but if there is some kind of interference that causes a different word to be more activated, a speech error may occur.

Lexical selection of a word also entails partial activation of the word’s phonological neighbors. Thus, in accessing *cat*, the words *hat*, *bat*, *cot* and so on are also activated. The activation of *cat* results in the phonemes /k/+/æ/+t/ being activated; the activation of *hat* results in the phonemes /h/+/æ/+t/ being activated, and so on. The end result is that the phoneme /k/ is in competition with /h/, /b/, and all the other possible initial phonemes. (The phonemes /æ/ and /t/ are also involved in competitions; we will ignore them here for the sake of simplicity, but the process is the same.) Baese-Berk and Goldrick (2009) noted that this competition is stronger between phonemes that share features—presumably, the strength of partial activation of neighbors is determined by the featural similarity of the target word to its neighbors, although their paper did not provide details. To inhibit these competitors, the target phoneme /k/ becomes more highly activated to ensure that it is selected. This activation then passes to the phonetic representation, where greater levels of activation lead to more extreme articulatory realizations.

Using this model, then, accessing a word like *cod* will cause activation of the neighbor *god*. Since /k/ and /g/ differ in only one feature ([±voice]), the competition is especially fierce, and /k/ needs a high level of activation to be selected. This high level of activation is passed to the phonetic representation, and consequently, the word
cod is produced with a relatively long onset VOT. In contrast, the word cop does not have a neighbor gop. No intense competition occurs, and the initial /k/ is passed to the phonetic level with a relatively low activation level. The word cop is pronounced with a relatively short onset VOT. When the word cod is presented in a context alongside god, the VOT enhancement is even greater, since god is already partially activated (due to being in the context) before the lexical access process begins. This explanation is how Baese-Berk and Goldrick (2009) account for their results.

According to this model, greater activation of a phoneme leads to a more extreme articulation of that phoneme. In the case of /k/, this extreme articulation has been claimed to result in a longer VOT, but it would also enhance other acoustic properties of the realization—for instance, a longer, louder burst phase and more extreme formant transitions. In the case of the VOT of cod being enhanced in order to distinguish it from god, then, we would still expect to see enhancements of formant transitions and other acoustic properties not related to the /k, g/ contrast. Similarly, in a situation where cod is contrasted with pod, the model predicts a VOT enhancement despite there being no VOT contrast between /k/ and /p/. The model could be partly falsified if these model predictions can be shown to be false.

Some evidence in favor of this prediction comes from experiments by Kirov and Wilson (2012; 2013), which showed that VOT enhancement occurs between pairs like cap and tap, even though these distinctions in place are not primarily cued by VOT contrasts. Further investigation of Baese-Berk and Goldrick’s (2009) results has suggested that the relevant lexical factor is not the presence or absence of a minimal pair in the word's neighborhood, but rather the size of the word's neighborhood as a whole that determines VOT production (Fox et al., 2015; Fricke et al., 2015). In particular, Fricke et al. (2015) proposed that a position-specific measure of neighborhood density, which
is able to distinguish onset stops from coda stops, is best-suited to accounting for these results.

Evidence of a contextually motivated reduction (or enhancement) that can be definitively distinguished from an activation account would be strong evidence against the model. However, because the model is vague with regards to how globally contextual information is realized as activation, is not clear how such a distinction would be made.

An alternative route to falsifying the model is related to its fundamental architecture. It is worth noting that this model of speech production is based on an oversimplified view of the production process. Unitary symbolic phonemes are concatenated—/k/+/æ/+/t/—and the concatenation is then passed to an SPE-style (Chomsky and Halle, 1968) phonetic implementation module to be translated to muscular motions that result in the physical actuation of speech. This ‘alphabetic’ view of phonology has a number of limitations (see e.g. Ladd, 2014; Munson et al., 2010). Work in articulatory phonology (Browman and Goldstein, 1992) and task dynamics (Saltzman and Munhall, 1989) has revealed that speech production cannot be easily accounted for with a simple mapping from abstract feature values to acoustic properties. Indeed, the targets of speech production have been proposed to be “multidimensional regions in auditory perceptual space” (Guenther et al., 1998, p622). The process of speech production itself has been proposed to be actuated by a feedforward control system (Guenther et al., 2006; Perkell, 2012), integrating online feedback with memorized gestures to create sounds appropriate to the physical context. Many contemporary psycholinguistic models of phonological planning, including the cascading activation, would need significant overhaul to be able to interface with the speech production models just mentioned.
1.2.3 The passive evolutionary perspective

Finally, the passive evolutionary perspective differs from the previous two perspectives in holding that no active force is responsible for predictability effects. Specifically, rather than communicative pressure or cognitive architecture producing these effects, phonetic reduction simply exists as a natural consequence of patterns of language acquisition and change over generations. Segments or words that are easy to perceive are generally perceived correctly, whereas segments or words that are difficult to perceive are only perceived correctly if they are sufficiently acoustically prominent. Over time, the sounds and words that are perceived correctly (i.e. easy words and acoustically prominent hard words) become the principal component of language; all other modes of production fall into disuse (Silverman 2012; see also Garrett and Johnson 2013). Silverman (2012, p147) expressed this position as follows (emphasis in original):

Successful speech propagates; unsuccessful speech does not. Confusing speech tokens may be misunderstood, and thus not pooled with the exemplars of the intended word, and so the system maintains its state of semantic clarity. Anti-homophony is thus not an active pressure for which there is an abundance of overt evidence. Rather, it is a passive result of the pressures that inherently act upon the interlocutionary process.

Implicit in this perspective is the assumption of an exemplar-style model whereby speech production proceeds as a process of selection of previously-experienced tokens. This mechanism is how correctly perceived sounds and words are able to propagate and misperceived sounds and words are not repeated. This first assumption is the ‘evolutionary’ part of the perspective. The other assumption is that reduced forms are easier and require less effort to produce than clear forms. Were this not the case, the
model would predict that ‘easy’ words can be pronounced either in a clear way or in a reduced way. This second assumption is shared with the listener-oriented perspective.

1.2.3.1 Evidence of language being optimized for communication

This partially-overlapping set of assumptions between the passive evolutionary account and the listener-oriented account means that much of the evidence in favor of the listener-oriented perspective can also be used in favor of the passive evolutionary perspective. Nevertheless, some of the evidence fits more neatly into an evolutionary perspective than into a listener-oriented account. For instance, work on homophone avoidance in historical change (Blevins and Wedel, 2009; Kaplan, 2011; Silverman, 2010; Tsui, 2015; Wedel et al., 2013a,b) has revealed that language change tends to avoid the creation of homophony, but it is not clear if this result comes from an active force that affects speakers day-to-day, or if it is a selectional process that operates over eras. Under the passive evolutionary account, the propagation of particular patterns of reduction in speech relies upon communication between individual language users. Communicative approaches to phonology have suggested that the phonological structure of the lexicon is optimized for efficient communication (Graff, 2012; Martin, 2007, 2011), and similar claims for optimal communication have been made for morphology (Caballero and Kapatsinski, 2015), pragmatics (Rohde et al., 2012), and kinship terms (Kemp and Regier, 2012).

Additional support for this evolutionary perspective comes from animal behavior research suggesting that non-human animal communication systems are structured to allow for maximal information transmission with minimal effort (see e.g. Bezerra et al. 2010; Semple et al. 2010, 2013 on primates and Luo et al. 2013 on bats). Ferrer-i-Cancho et al. (2013) explicitly argued that all communication systems, including
human language, are governed by basic distributional properties which enhance efficiency of coding. These findings are relevant for the present discussion to the extent that human language evolved from similar primitive communication systems (Hurford, 2007), because any evolved communication system must be shaped in an efficient way: if communication is not successful, any attempt at communication is a waste of effort and thus will be selected against, causing the communication system to collapse and disappear (Ackley and Littman, 1994). Alternatively, if the production of the communicative signal requires too much effort, less energy is available for reproduction or survival, and communication will again be selected against. Thus, for a communication system to persist, successful communication with the lowest possible energy expenditure is necessary (see Ferrer-i-Cancho and Elvevåg, 2010; Ferrer-i-Cancho and Moscoso del Prado, 2011, for statistical approaches to this reasoning). From this view, the observed ‘predictability effects’ are a necessary consequence of natural selection. No appeal to cognitive or psychological mechanisms is needed. If all communication systems necessarily evolve predictability effects as a simple consequence of being a communication system, then there is no phenomenon to explain, any more than we need to explain why our legs reach the ground when we stand up.

1.2.3.2 Exemplar dynamics

One of the few explicit formulations of the passive evolutionary perspective comes from Janet Pierrehumbert's work on exemplar-based phonology (Pierrehumbert, 2001a,b, 2002, 2003a,b). This description involves an exemplar model (Goldinger, 1998; Johnson, 1997; Tenpenny, 1995) in which each perceived word token has its own representation in a perceptual cloud (see also Blevins and Wedel, 2009; Tupper, 2014; Wedel, 2006, for refinements and extensions of these mechanisms). In Pierrehumbert’s (2002) model, predictability effects emerge as a simple consequence of the acquisition process.
Consider the scenario where a token of a high-frequency word is uttered. Even if the word is not particularly perceptually clear, the listener can guess the word’s identity with relative ease due to its frequency. When the word is identified, the token is added to the listener’s exemplar cloud and becomes part of that word’s representation. However, when a token of a low-frequency word is uttered, the listener cannot guess the word’s identity as easily, due to its low frequency, and the token therefore needs to be more acoustically prominent than the high-frequency word in order for its identity to be ascertained correctly. When the word is not correctly identified, the token is not added to the listener’s exemplar cloud and does not become part of the target word’s representation. Thus, the low-frequency word token will only be added to the exemplar space if it is sufficiently acoustically prominent (Tupper, 2014). Over time, then, the exemplar space will contain acoustically prominent low-frequency words, and both prominent and non-prominent high-frequency words. In speech production, the talker selects a token at random from the exemplar space of the target word (see Pierrehumbert, 2001a, 2002, for mathematical details of the implementation). High-frequency words will tend to be reduced in production relative to low-frequency words because their exemplar clouds contain both reduced and unreduced variants, whereas the exemplar clouds of the low-frequency words contain primarily unreduced variants, leading to unreduced productions of these targets. Within a speech community, this behavior becomes a positive feedback loop leading to clear productions of low-frequency words and reduced productions of high-frequency words.

However, in order for this model to account for more than simple frequency effects, the exemplar representations must be made much richer to be able to distinguish different kinds of discourse contexts that lead some words or phrases to be more predictable. Such a richer representation also needs to be accompanied by a corresponding increase in abstraction in the production process, lest this model of language
be reduced to an imitative parrot with no generalization or generativity. These issues are discussed in greater depth in Chapter 5.

1.3 Research questions

The goals of this dissertation are to investigate

- the role of theory of mind in the production of predictability effects (Chapter 2);
- the extent to which unreduced speech is perceptually beneficial relative to reduced speech, and the extent to which theory of mind modulates individual differences in this regard (Chapter 3); and
- whether predictability effects can be ascribed to a single cognitive source, or if multiple factors are implicated (Chapter 4).

The remainder of this chapter is devoted to exploration of theory of mind and its methods of assessment in neurotypical adults.

1.4 Theory of mind assessment

Lack of theory of mind has been proposed as one of the primary cognitive deficits observed in people with autism spectrum disorders. This section defines theory of mind and describes how it is assessed in neurotypical populations. Included is an overview of theory of mind assessments and what they are claimed to measure.

Theory of mind (henceforth ToM) is the ability to attribute mental states to others; if someone possesses a ToM, it means “that the individual imputes mental states to others” (Premack and Woodruff, 1978, p515). This definition naturally includes a wide range of potential states, including those of knowledge, intention, purpose, desire, emotion, and so on, for an individual to attribute to others. For present purposes,
we will use Premack and Woodruff's above definition, necessarily excluding some definitions which would have ‘empathy’ rather than ToM be responsible for emotional mental states (see Decety and Jackson, 2004).2

ToM as thus defined is not directly observable, and various methods have been created to assess whether or not an individual possesses ToM. The majority of ToM assessment methods are designed for use with children, either as a diagnostic aid or as a research method to compare different clinical populations, but some work with adults does exist.

1.4.1 False belief

One of the oldest methods of assessment of ToM is the false belief task, developed by Wimmer and Perner (1983) and elaborated as the now widely-known ‘Sally–Anne test’ by Baron-Cohen et al. (1985). The task involves two dolls called Sally and Anne, which are controlled by the experimenter (see Figure 1.2). Sally places a marble into a basket, and then leaves. In her absence, Anne takes the marble and hides it in the box. Sally then returns. The child is asked where Sally will look for the marble. If the child points to the basket, it demonstrates that the child is able to model Sally’s false belief, despite the child’s knowledge of the actual position of the marble. If the child points to the box, the child demonstrates their lack of ability to account for the beliefs of others. Three control questions are also asked—at the beginning, the child is asked the doll’s names; and at the end, the child is asked where the marble was originally, and where it is now. If these control questions are answered correctly then drawing

---

2Although the concept of ToM is usually regarded as originating with Premack and Woodruff (1978), the related concepts of metarepresentation and mentalizing have been part of biological thinking for many decades; see chapter 12 of Hobhouse (1901) for an early treatment of the subject.
Figure 1.2: The Sally–Anne false belief test. Figure from Baron-Cohen et al. (1985).

conclusions from the answer to the belief question is warranted. Variants of the Sally–Anne test have also been performed with gaze—an implicit behavioral measure—as the dependent variable, rather than verbal responses (Garnham and Ruffman, 2001; Ruffman et al., 2001). The location the child looks to while considering the question reveals their anticipation of Sally's action.

Another test used to assess ability to comprehend false beliefs is the so-called ‘Smarties task’ (Gopnik and Astington, 1988; Hogrefe et al., 1986; Perner et al., 1987, among others). In this task, a child is shown a tube of Nestlé Smarties, an easily identifiable candy.³ The child is asked what they think is inside the tube (the answer

---

³Nestlé Smarties are sugar-coated chocolates popular in the UK, Canada, Europe, and elsewhere, and are sold in a small cardboard tube. They are not to be confused with the US-based Smarties Candy Company’s fruit flavored tablet ‘Smarny’ of the same name, which are instead sold in plastic wrappers.
is invariably “Smarties”), and then the child is shown that the tube actually contains a pencil. After answering control questions about the current content of the tube and the child’s previous belief, the child is asked what a friend will say when they are brought in and asked what is in the tube. The answer ‘Smarties’ suggests that the child is able to model their friend’s mental state of not knowing that the tube contains a pencil; the answer ‘a pencil’ suggests that the child has failed to do so. Although this task is similar to the Sally–Anne test, the Smarties task attempts to have the participants understand the false belief of another through direct experience of their own false belief, rather than pure reasoning alone (in this connection, see comments by Davis, 1978).

Bloom and German (2000) argue that the false belief test is inappropriate as a measure of theory of mind; possessing ToM, they claim, is neither necessary nor sufficient to pass a false belief test. The false belief task is a demanding task for various reasons, and it has been found that 3-year-old children who fail a false belief task also fail at similarly complex tasks that do not involve false beliefs (Zaitchik, 1991). There is also evidence that children who do not pass a false belief test still nevertheless possess ToM: 2-year-old children (who usually fail false belief tests) have been shown to use different request strategies depending on whether or not the parent was able to observe the toy being hidden (O’Neill, 1996, 2005), a clear indicator that the child is aware that the parent has an understanding of the world around them that is separate from that of the child’s. Bloom and German (2000) cited a number of related studies demonstrating plausible ToM skills in young children who fail a false belief test. They concluded that the false belief test is a difficult task and one that only reveals some aspects of an individual’s ToM abilities. More recently, Yott and Poulin-Dubois (2012) presented evidence suggesting that 18-month-old infants possess an implicit understanding of false belief, despite their lacking the linguistic capacity to respond to a traditional false belief test.
1.4.2 Other aspects of theory of mind

Despite the focus on false belief tests, there are many other aspects of ToM that have been addressed in the literature (see Baron-Cohen, 2000; Wellman and Lagattuta, 2000, for reviews). One of these aspects is the difference between the physical world and the mental world. Suppose a child is listening to a story where one character is thinking about a dog, while another is holding a dog. Both characters are involved in an experience with a dog—one mental, one physical. The child is then asked which character is able to stroke the dog—children with a developed theory of mind are able to correctly identify that only the child involved in the physical experience with the dog is able to stroke it (Wellman and Estes, 1986). In a similar task, the child is shown a candle shaped like an apple, and asked about the identity of the object. Flavell et al. (1986) reported that three- and four-year-olds are able to explain that the object looks like an apple, but in reality it’s a candle.

A consequence of the physical–mental distinction is the ability to understand that emotions have mental causes in addition to physical causes. For instance, someone can be sad because they have hurt themselves—a physical cause—or they can be sad because they didn’t get what they wanted—a mental cause. An indirect method of assessing knowledge of the difference between the mental and physical is how a child uses and categorizes different words. Some words refer to mental events (such as think, know, pretend), while some refer to physical events (such as jump, eat, dance). Children with ToM skills use these mental words more often and are able to classify words into ‘mental’ and ‘physical’ categories successfully (Baron-Cohen et al., 1986, 1994; Tager-Flusberg, 1992).

Another language-related aspect of ToM is the ability to understand figurative speech, such as metaphor, idioms, sarcasm, and irony. Without an understanding of
the mental states of others, sarcasm is impossible to detect, and metaphors and idioms become extremely opaque. The detection and practice of deception also crucially hinges upon having a mental representation of another's state of mind. Additionally, being able to impute the mental state of another implies that one can impute the mental state of oneself—i.e. metacognition, the ability to monitor and be aware of one's own intentions and state of mind (see Lombardo and Baron-Cohen, 2011). Finally, the ability to infer desire or attention from observing another’s gaze has been proposed as a metric of ToM (Schneider et al., 2012a).

ToM, then, manifests itself as a cluster of related mind-reading abilities. As such, it is perhaps simplistic to speak of an individual as either ‘having’ or ‘not having’ ToM; rather, the abilities are expressed in gradient degrees (Yirmiya et al., 1998). The understanding of ToM abilities being on a continuum of aptitude is reflected in many formal ToM assessment methods (see e.g. Happé, 1994, and others in the following subsection). This view is supported by evidence suggesting that ToM follows a consistent developmental pattern in typically-developing children (Wellman and Liu, 2004). Research has also shown that even adults show variance and individual differences in their extent and application of ToM skills. Some adults fail to apply ToM abilities in communicative situations (Keysar et al., 2003); and there is growing evidence that ToM is not an automatic ability and thus requires effort and suffers when the individual is under a high cognitive load (Lin et al., 2010; Schneider et al., 2012b), despite the fact that ToM is not related to intelligence (Rajkumar et al., 2008). Uekermann et al. (2007) reported ToM deficits in individuals diagnosed with alcoholism, possibly due to brain damage from alcohol abuse (see also Rowe et al., 2001; Uekermann and Daum, 2008); Wolkenstein et al. (2011) identified ToM deficits in individuals with acute depression. Non-pathological individual differences in ToM could also arise as a result of variation in personality traits (McIlwain, 2003).
The ability to impute mental states to others entails several direct and indirect capabilities. Understanding others’ false beliefs and emotions, figurative speech, and being able to reason about why another person believes, feels, or talks in a particular way is a cornerstone of theory of mind. Applying the same reasoning to oneself is a necessary first step. Additionally, ToM is revealed implicitly through eye gaze, the understanding of the physical–mental distinction, and the use of language. The next subsection discusses several formal assessments of aspects of theory of mind appropriate for use with adults, besides the Sally-Ann test.

1.4.3 Alternatives to false belief tasks

Fortunately for the ToM researcher, there exist several alternatives to the false-belief task. I will discuss four of them here: the ‘Reading the Mind in the Eyes’ test; the ‘Strange Stories’ task; the ‘Social Attribution Task’; and the ‘TOM Test’.

In the Reading the Mind in the Eyes test, the participant is presented with a photograph of the eye region of the face (see Figure 1.3 for an example), and asked

Figure 1.3: Example stimulus from the ‘Eyes Task’. Is this person jealous, panicked, arrogant, or hateful? From Baron-Cohen et al. (2001a).
which of four adjectives best describes what the person in the photograph is feeling (Baron-Cohen et al., 1997a, 2001a, 1997b). This test assesses ability to infer complex emotional states from simple pictures, and thus serves as an indirect measure of theory of mind. The test has been shown to be a reliable and stable measure of ToM in non-clinical adult populations (Fernández-Abascal et al., 2013; Vellante et al., 2013).

In the Strange Stories task, the participant reads short, paragraph-length stories and answers a question about the motivation and relationships between the characters of a story (Happé, 1994; Jolliffe and Baron-Cohen, 1999; White et al., 2009). For example, subjects may be asked why a particular character told a lie or did some unusual action. These stories are scored based on whether or not the answer was correct and whether or not the answer made reference to the mental state of others. An example story is provided in (5), its associated question in (6), and the scoring rubric in (7), from White et al. (2009).

(5) Simon is a big liar. Simon’s brother Jim knows this, he knows that Simon never tells the truth! Now yesterday Simon stole Jim’s ping-pong paddle, and Jim knows Simon has hidden it somewhere, though he can’t find it. He’s very cross. So he finds Simon and he says, “Where is my ping-pong paddle? You must have hidden it either in the cupboard or under your bed, because I’ve looked everywhere else. Where is it, in the cupboard or under your bed”? Simon tells him the paddle is under his bed.

(6) Q: Why will Jim look in the cupboard for the paddle?
This test distinguishes ToM ability on the basis of how often participants make reference to the mental states of others versus reference to the physical world.

Klin (2000) presented the Social Attribution Task, where participants watch a silent video from Heider and Simmel (1944). The video has a ‘cast’ of four geometric shapes—a large triangle, a small triangle, a small disc, and a large rectangle with a door-like opening—see Figure 1.4. The shapes move in different directions at different speeds throughout the film, and it is possible to describe the shapes as animate characters with desires and motivations. Heider and Simmel (1944) reported that only 3 of their 114 participants did not interpret the movements of the shapes as human (or otherwise animate) actions; the majority of participants were quick to invent explanatory stories, involving detailed dramatis personae and themes of love, betrayal, and even abuse. The film is little more than a minute long.

In the Social Attribution task, after watching the film twice, the participant narrates the main events of the film. Following that, the film is split into six meaningful scenes which are shown to the participant who narrates each one in turn. Next, the participant is told (or reinforced) to consider the shapes as people, and asked what kind of a person each of the three filled shapes is. Finally, the participant is asked seven questions about specific events, e.g. “why did the big triangle break the house?” Note that some of these questions explicitly name objects or events in the film (e.g. ‘house’ and ‘fighting’). This procedure generates 1 main narrative + 6 scene narratives + 3 character narratives + 7 question narratives = 17 total narratives per participant for
The experiments on the perception of the behavior of others here reported are in method and purpose different from the investigations mentioned. In the first place, instead of presenting faces with the exclusion of the situation, we have presented situations and activities without the face. Secondly, our aim has not been to determine the correctness of the response but instead the dependence of the response on stimulus-configurations.

Figure 1.4: Example of the shapes displayed in the film. From Heider and Simmel (1944). An internet search for 'heider simmel 1944 video' provides multiple resources to view the full video.

analysis. Each narrative is scored on seven difference indices (pertinence, salience, ToM cognitive, ToM affective, animation, person, problem-solving) each of which assesses an aspect of ToM and social engagement with the video. This test measures the degree to which participants are able to impute mental states to geometric objects, and their perception of social interactions between said objects.

The TOM-Test is described by Muris et al. (1999). In the test, the participant is shown a number of vignettes, drawings, and stories and asked questions about each of them in turn. Each item is allocated to one of three subscales: (1) “precursors” to ToM, such as emotion recognition; (2) “first manifestations” of a full ToM, such as false belief; and (3) “more advanced” ToM, including humor and understanding second-order beliefs. Figure 1.5 shows one such picture, which is accompanied by the instructions and questions shown in (8).
Figure 1.5: Example drawing from Muris et al. (1999) to accompany the questions in (8).

(8) Instruction: Take a look at this picture. (Figure 1.5)
Question 1: What has happened? Can you tell something about it? (subscale 1)
Question 2: Who in this picture is afraid? (subscale 1)
Question 3: Why is this person afraid? (subscale 2)
Question 4: Who in this picture is happy? (subscale 1)
Question 5: Why is this person happy? (subscale 2)
Question 6: Who in this picture is sad? (subscale 1)
Question 7: Why is this person sad? (subscale 2)

Question 8: Who in this picture is angry? (subscale 1)

Question 9: Why is this person angry? (subscale 2)

The third subscale is tested in the story and questions shown in (9).

(9) Instruction: I will read you a short story. Listen carefully.

Story: It is summer. Will and Mike have their holidays. They go out for a bicycle ride. Suddenly, there is a downpour and they have to shelter in a bus station. There are two men in the bus station who also shelter from the rain. One of the men remarks: “Wow, we have nice weather today!”

Question 1: What does the man mean?
Question 2: Is it true what the man says?
Question 3: Why does the man say: “Wow, we have nice weather today!”

Each question is scored simply as correct or incorrect. This test examines a wide range of ToM abilities, but does not differentiate them.

1.4.4 The Autism-spectrum Quotient

A common method for assessing autistic traits in adults of normal intelligence is the Autism-spectrum Quotient (AQ; Baron-Cohen et al., 2001b). The task takes the form of a short self-reported questionnaire, providing statements such as “I am good at social chit-chat” to which the test-taker responds on a four-point agree/disagree scale. Due to its ease of administration and scoring and a high test–retest reliability, the AQ has seen wide use in developmental research since its introduction (but cf. Bishop and Seltzer, 2012; Nishiyama et al., 2014).
The AQ was designed to be used to measure autistic traits, not theory of mind. However, to the extent that the central cognitive deficit of autism is a lack of ToM (Baron-Cohen, 1995; Baron-Cohen et al., 1985), the AQ can be used as a measure of ToM. Indeed, in Senju et al.’s (2009) study of anticipatory eye movements in people with Asperger syndrome during a false belief task, the groups of participants (one group of people with Asperger syndrome and another group of neurotypical adults) were matched as much as possible in terms of IQ and verbal ability. Participants also underwent a battery of ToM tests (including several false belief tests and the Strange Stories task) and completed the AQ questionnaire. The only significant difference between the groups was in AQ score—the Asperger syndrome group had much higher AQ scores than the neurotypical group—i.e. no differences between groups in the ToM tasks was observed. Additionally, the experimenters reported finding significant results in terms of eye movements, revealing that the Asperger syndrome group has a ToM deficit relative to the neurotypical group. Since AQ was the only metric differentiating the groups, then, the AQ scores serve as a measure of ToM, capturing that which the other tests were unable to adequately assess.

Of these tasks, the Reading the Mind in the Eyes test, the Strange Stories task, and the Autism-spectrum Quotient were used in Chapter 2 to assess ToM variation in neurotypical adults. These tasks were chosen on the basis of being relatively short, easy to administer via computer with minimal input from an experimenter, and relatively

---

4 Asperger syndrome is an autism spectrum disorder. There is disagreement about the extent to which Asperger syndrome is the same nosological entity as high-functioning autism (see Macintosh and Dissanayake, 2004; Tsai, 2013, for review). Consequently, the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013), which offers diagnostic criteria for mental disorders, does not list Asperger syndrome as a possible diagnosis. Instead, the symptoms that led to a diagnosis of Asperger syndrome in previous editions of the DSM (e.g. American Psychiatric Association, 1994) now lead to a diagnosis of high-functioning autism. This change has been somewhat controversial (see e.g. Mahjouri and Lord, 2012).
straightforward to score. Experiment 1a includes some analysis of how effective these measures were at quantifying ToM in the participant sample.

1.5 Overview of thesis

The rest of this dissertation presents nine experiments which investigate the ‘listener-oriented’ account of predictability-based phonetic reduction. Chapter 2 presents the results of a series of experiments designed to test the prediction of the listener-oriented account that relates to speech production and theory of mind, and Chapter 3 tests the assumption of the account relating to the perception of phonetically reduced speech. Chapter 4 involves corpus analysis and critically analyzes the concept that phonetic reduction is a single phenomenon. Chapter 5 concludes the thesis with comparison of other models of phonetic reduction and evaluates how well they are able to account for the present data. A speculative model is outlined, and ideas for future work are proposed.
Chapter 2

Theory of mind and phonetic reduction

2.1 Introduction

Phonetic reduction is pervasive in natural speech. Many instances of reduction can be linked to linguistic predictability, a phenomenon known as ‘predictability-based reduction’. Repeated from Chapter 1, the working definition of ‘predictability-based reduction’ is any phenomenon of reduction of some spoken language element (but especially a phenomenon of phonetic reduction that affects the magnitude of a physical property such as duration) which is mediated by or correlated with some observable property of the predictability of that element. This chapter examines the relationship between predictability-based reduction and theory of mind; such a relationship is hypothesized to exist by listener-oriented accounts of predictability-based reduction.

2.1.1 Predictability-based reduction

This chapter details three experiments which investigate interactions among theory of mind and various types of phonetic reduction. Five kinds of phonetic reduction are investigated: semantic predictability-based reduction; second mention reduction; word-specific VOT reduction; reduction due to neighborhood density; and reduction due to frequency.
“Semantic predictability-based reduction” refers to phonetic reduction which can be attributed to the word being semantically predictable in context. For example, the last word in (1) is more predictable given the preceding context than is the last word in (2), and Lieberman (1963) found that productions of this highly predictable token of ‘nine’ was less intelligible when presented in isolation relative to the less predictable version. In other words, the more predictable word was phonetically reduced. Semantic predictability-based reduction has been shown to affect both durations and vowel spectra for a number of different definitions of predictability (Aylett and Turk, 2006; Bell et al., 2009; Charles-Luce, 1993; Clopper and Pierrehumbert, 2008; Engelhardt and Ferreira, 2014; Gahl and Garnsey, 2004; Hunnicutt, 1985, 1987; Jurafsky et al., 2001; Lieberman, 1963; Moore-Cantwell, 2013; Pate and Goldwater, 2011; Pluymaekers et al., 2005a; Scarborough, 2010; Tily and Kuperman, 2012).

(1) A stitch in time saves nine.

(2) The number that you will hear is nine.

“Second mention reduction” refers to a word’s being reduced upon its second or subsequent use in a discourse (Fowler and Housum, 1987). This effect has been documented in a variety of studies, both corpus-based and lab-based (Baker and Bradlow, 2009; Bard and Anderson, 1994; Bard et al., 2000, 1989; Burdin and Clopper, 2015; Fowler, 1988; Fowler et al., 1997; Galati and Brennan, 2010; Kahn and Arnold, 2012, 2015; Kaiser et al., 2011; Lam and Watson, 2010, 2014; Pate and Goldwater, 2011; Sasisekaran and Munson, 2012; Shields and Balota, 1991; Turnbull, 2015; Vajrabhaya and Kapatsinski, 2011). The connection to predictability is that previously mentioned (discourse given) items are more predictable than previously unmentioned items, since they are established in the discourse context and more likely to be referred to than other items.
“Word-specific VOT effects” are a sub-class of neighborhood density effects (discussed below), and were first documented by Baese-Berk and Goldrick (2009) (but see also Goldinger and Summers, 1989). Their study examined the production of words with an initial voiceless stop. In particular, they compared words like cod, which have a minimal-pair neighbor god, with words like cop, which do not have a minimal pair neighbor (*gop). Their results showed a difference of approximately 5ms between the VOTs of the two sets of words, with cod-words having the longer VOT cue. Additionally, VOT was further enhanced by another 5ms when the minimal pair contrast between cod and god was contextually important to the experimental trial. They attributed these results to the existence of the minimal pair neighbor in the lexicon boosting the VOT of the target word, serving to make the target word more distinct from its voiced neighbor. Alternatively, this effect can be thought of as the lack of a minimal pair neighbor causing the VOT of the target to become reduced. This effect has been replicated in a number of studies (Bullock-Rest et al., 2013; Fox et al., 2015; Goldrick et al., 2013; Kirov and Wilson, 2012; Peramunage et al., 2011), although the source of the effect has been disputed. Fox et al. (2015) claimed that the effects were due to general phonological neighborhood density, rather than effects of specific minimal pairs. A conceptually similar effect was observed in Korean by Holliday and Turnbull (2015): word initial lax obstruents in high-density words tended to have lower F0 values than those in low-density words. In Korean, low F0 is an acoustic cue differentiating the lax obstruents from the tense and aspirated series; this F0 lowering can thus be regarded as cue enhancement similar to the English VOT effects. Fricke et al. (2015), through a re-analysis of Baese-Berk and Goldrick’s (2009) data set, argued that a position-specific measure of neighborhood density provided the most effective predictions of the results, echoing other calls in the literature for more nuanced approaches to lexical neighborhoods (Clopper and Tamati, 2014).
Phonological neighborhood density refers to the number of “neighbors” a word has. Neighbors are usually defined as words which differ from the target word by one phoneme insertion, deletion, or substitution (Pisoni et al., 1985). By this metric, the neighbors of feed include fee (by deletion), field (by insertion), food and seed (by substitution). The relation of neighborhood is symmetric (if X is a neighbor of Y then Y is necessarily a neighbor of X) but not transitive (fee and field are not neighbors). Many studies of neighborhood density attempt to account for the neighbors’ frequency, under the assumption that a more frequent neighbor has more ‘influence’ within a neighborhood than a low-frequency neighbor. This adjustment has been achieved by taking the summed frequency of all neighbors (e.g. Luce and Pisoni, 1998), or the target word frequency relative to the summed frequency of all neighbors (e.g. Scarborough, 2004, 2010, 2013), among other measures. Several studies have noted that words with more neighbors tend to have more disperse vowel productions than words with fewer neighbors (Clopper and Tamati, 2014; Munson, 2007; Munson and Solomon, 2004; Scarborough, 2004, 2010; Watson and Munson, 2007, 2008; Wright, 2004). Effects on duration have also been observed, with higher-density words and vowels tending to be longer than lower-density words and vowels (Burdin and Clopper, 2015; Burdin et al., 2014a,b; Kryuchkova and Tucker, 2012; Scarborough, 2010). However, the literature is somewhat inconsistent; among the papers that report vowel-specific effects, there is disagreement on which vowels are affected by neighborhood density and which are not. Table 2.1 summarizes findings from Wright (2004), Watson and Munson (2007), and Scarborough (2010), indicating with a checkmark (✓) where a significant effect was observed and a cross (×) where it was not. As can be seen, of the eight vowels, the only unanimous agreement is that neighborhood density influences /æ/ but not /ɛ/ (and cf. Clopper and Tamati, 2014, which casts further doubt on the precarious status quo). An additional area of inconsistency was highlighted by Gahl et al. (2012),
who claimed that higher neighborhood density leads to shorter and less disperse vowels, and that previous studies had failed to control for numerous segmental and contextual influences on vowel production that affect formant measures (see also Gahl, 2015).

Finally, phonetic reduction is related to frequency, such that high-frequency words tend to be more reduced than low-frequency words (Bybee, 2002, 2006; Gahl, 2008; Jurafsky et al., 2001). Frequency can be directly related to predictability as unigram probability, or a priori probability of a word's occurrence. These five kinds of reduction—semantic predictability-based, second mention, word-specific VOT, neighborhood density, and frequency—are examined in this chapter.

### 2.1.2 Theoretical perspectives

The 'listener-oriented' accounts of predictability-based reduction hold that there are direct and active processes that serve to enhance the perceptibility of speech in low-predictability conditions. The inhibition of these processes, the accounts further posit, is the cause of phonetic reduction in speech. A classic formulation of this view is Björn Lindblom's 'Hyper- and Hypo-Articulation Theory' (H&H theory), where the talker is

<table>
<thead>
<tr>
<th></th>
<th>i</th>
<th>e</th>
<th>e</th>
<th>ə</th>
<th>a</th>
<th>o</th>
<th>u</th>
<th>ʌ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wright (2004)</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Watson and Munson (2007)</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Scarborough (2010)</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2.1: Summary of reported effects of neighborhood density on vowel dispersion from three studies. A checkmark signifies a significant effect was observed, a cross signifies the lack of an effect. Empty cells indicate the study did not consider that column's vowel.
assumed to have “tacit awareness of the listener’s access to sources of information independent of the signal and his [the listener’s] judgement of the short-term demands for explicit signal information” (Lindblom, 1990, p403). In other words, the talker knows which aspects of the signal are highly predictable due to context or language structure; these forms are therefore somewhat redundant and the talker can reduce them (i.e. ‘hypoarticulate’). The talker also knows which aspects of the signal the listener may misperceive; the talker can emphasize these forms (i.e. ‘hyperarticulate’). The whole act of speech is a balance between conservation of effort (causing hypoarticulation) and attending to listener need (causing hyperarticulation).

Other proposals exist which, although not exactly the same, have a similar conceptual basis—predictability effects exist as a means to ensure smooth and uninterrupted communication between interlocutors; speech is an adaptive and goal-oriented process. These proposals include the Smooth Signal Redundancy Hypothesis (Aylett, 2000; Aylett and Turk, 2004, 2006; Turk, 2010), and those of Flemming (2010); Fox Tree and Clark (1997); and Levy and Jaeger (2007). In phonology, several phonological processes have been explained as being partially motivated by perceptual salience, which points to the role of audience design in determining language structure; see e.g. Jun (1995, 2004, 2010); Steriade (2008); Silverman (2012, ch. 8); and Hume and Johnson (2001). Additionally, the audience design reasoning has been applied to linguistic predictability effects outside of phonetics, see e.g. Galati and Brennan (2010) and Horton and Gerrig (2002) for discourse structure and Jaeger (2010), Kraljic and Brennan (2005), Levy and Jaeger (2007); and Lockridge and Brennan (2002) for syntax.

A strong interpretation of the listener-oriented account holds that the talker must be able to create, maintain, and update a detailed mental representation of their interlocutor’s knowledge, beliefs, intentions, desires, and emotions in real time—that
is, the talker must possess a well-developed theory of mind, the ability to impute complex mental states to others. From this interpretation it follows that individual variation in theory of mind ability is linked to the extent and application of predictability-based reduction. This prediction is tested in this chapter, by examining correlations between scores in theory of mind tasks and the size of acoustic reduction in language production.

It follows from a strong interpretation of the listener-oriented account that the talker must employ a variety of ToM skills in order to be able to communicate efficiently (i.e., to reduce and enhance in the appropriate places). According to this reasoning, without full command of these ToM abilities, the talker is not able to know when to reduce and when to enhance. Regardless of one’s theory of how the speech production process unfolds, the listener-oriented account requires that the talker’s knowledge of the listener exerts some influence upon speech. Figure 2.1 shows a schematized and

Figure 2.1: Simplified boxology of influence of knowledge of one’s interlocutor on speech production. See text for details.
simplified boxology of this process. The speech production process has some higher-
level processes (such as syntax, phonological planning, and other details) which feed
into some process of phonetic planning. From the phonetic planning, speech is pro-
duced. The figure is deliberately agnostic as to whether this ‘phonetic planning’ is
in terms of symbolic phonological units, articulatory kinematics, gestural scoring, or
some other such process. The precise details are irrelevant. In any case, according to a
listener-oriented account, the knowledge of one’s interlocutor intervenes at some point
to reduce or enhance certain phrases, words, syllables, phones, or some other unit of
planning. If it is known that the interlocutor is likely to find some unit predictable,
it can be reduced. Unpredictable units can be enhanced. Thus this subsystem bears
some influence onto the speech production process. In an extreme case, if there is no
knowledge of the interlocutor, that is, the talker does not possess a theory of mind,
then there is no influence and thus no observed predictability effects. However, all
neurotypical individuals (are assumed to) possess some degree of ToM; the variation is
in how easily they are able to use it. In the case of the present boxology, then, for talk-
ers with a weaker ToM, the knowledge of the interlocutor and its relevance for speech
communication will be delayed in its calculations. Such a delay will lead to a smaller
degree of enhancement or reduction, since the subsystem has had less time to exert its
influence before the talker moves on to planning the next word.¹ For talkers with a
stronger ToM, the calculations of what to enhance and reduce will be that much faster
and easier, and thus the effects are predicted to be larger. The notion that taking into
account the perspective of others, a ToM skill, is an effortful process which takes time
is particularly prominent in the work of Boaz Keysar (e.g. Horton and Keysar, 1996;

¹For convenience of exposition, the present toy model is assuming that speech production proceeds
in an orderly word-by-word manner. This simplifying assumption is incorrect—speech planning is much
more holistic. However, as mentioned, the present predictions hold for a listener-based account regard-
less of one’s theory of speech planning.
Keysar et al., 2000, 2003; Lin et al., 2010). Such a perspective implies that variation in use of ToM can be expected.

Weaker interpretations of the listener-oriented account exist, which do not necessarily require interlocutor-specific knowledge but instead rely on ‘generic listener’ models. Turk (2010, p230) noted:

Crucially, the Smooth Signal Redundancy proposal does not require that the speaker necessarily take the listeners into account during the online speaking process. The speaker’s language redundancy computation can be made on the basis of his or her own language experience. While not necessarily optimal for the listener, this type of language redundancy computation may represent a reasonable approximation to the language redundancy of the listener. Information about the listeners’ knowledge or experience can be incorporated in the computation, but doesn’t have to be.

However, it is difficult to reconcile this approach with existing literature on audience design in collaborative tasks. In classic tasks in the action-language tradition, two interlocutors must work together to arrange a set of tangram figures (Clark and Wilkes-Gibbs, 1986). During the course of the task, the interlocutors develop ‘conceptual pacts’ whereby the name of a particular figure is conventionalized, such as “the ice-skater” (Brennan and Clark, 1996). A similar study by Isaacs and Clark (1987) used New York City landmarks instead of tangrams, and had dyads of two experts (native New Yorkers) or an expert and a novice. They found that interlocutors are extremely adept at quickly determining the knowledge level of their dialogue partner, and tailoring their productions accordingly. These findings illustrate that talkers consistently employ listener-modeling in their choice of referring expression. The weak form of the listener-oriented account quoted above holds that talkers can do listener-modeling
when needed, but that it is not required. This version of the theory does not make explicit the conditions under which we expect talkers to use listener-specific knowledge, which means that testing the validity of this account is difficult.

2.1.3 Theory of mind

Theory of mind (henceforth ToM) is the ability to attribute mental states to others; if someone possesses a ToM, it means “that the individual imputes mental states to others” (Premack and Woodruff, 1978, p515). This definition naturally includes a wide range of potential states, including those of knowledge, intention, purpose, desire, emotion, and so on, for an individual to attribute to others. Rather than a binary “you have it or you don’t”, ToM is best considered as a cluster of related mind-reading abilities, including the ability to comprehend false beliefs, the ability to make distinctions between the physical and mental worlds (including physical vs. mental causality), and the ability to understand figurative language (see Baron-Cohen, 2000; Wellman and Lagattuta, 2000, for reviews). Baron-Cohen et al. (1985) proposed that a ToM deficit is a “crucial component” (p37) of the social impairment symptoms of autism spectrum conditions. Although the centrality of this deficit has been challenged (e.g. Bennett et al., 2013; Boucher, 2012; Happé and Frith, 2006; Mottron et al., 2006; Tager-Flusberg, 2003), there is ample evidence that such a deficit exists (but cf. Begeer et al., 2010).

Although early work focused on individuals ‘having’ or ‘not having’ ToM (e.g. Baron-Cohen et al., 1985; Premack and Woodruff, 1978), ToM abilities are expressed in gradient degrees and ToM is best understood as a continuum of aptitude (Happé, 1994; Yirmiya et al., 1998). Research has shown that neurotypical adults show variability and individual differences in their extent and application of ToM skills: some adults fail to apply ToM abilities in communicative situations (Keysar et al., 2003); and there is
growing evidence that ToM is not an automatic process and thus requires effort and suffers when the individual is under a high cognitive load (Lin et al., 2010; Schneider et al., 2012b), despite the fact that ToM skills are not related to intelligence (Rajkumar et al., 2008). Non-pathological individual differences in ToM can also arise as a result of differences in personality traits (McIlwain, 2003) and autistic traits, which are normally distributed throughout the neurotypical population (Caldwell-Harris and Jordan, 2014; Constantino and Todd, 2003).

Although it is common to speak of ToM as a unitary cognitive process, there is evidence that these abilities arise from an interplay of different mechanisms. For instance, Tager-Flusberg and Sullivan (2000) have argued that the social-cognitive and social-perceptual aspects of ToM are separate, and Apperly and Butterfill (2009) have argued that there is an automatic and implicit ToM mechanism which operates independently of (and develops far earlier than) a conscious and explicit ToM mechanism (see also Cohen et al., 2015). A full examination of these issues, although interesting, is beyond the scope of this chapter. For the purposes of this study, ToM is treated as having a single dimension and is parameterized through performance in standard assessment methods, discussed below.

2.1.4 ToM assessment

There exists a wide literature on methods to assess extent of ToM in individuals (see e.g. Baron-Cohen et al., 1997a, 1985, 1986, 1994, 2001a, 1997b, 2001b; Blijd-Hoogewys et al., 2008; Flavell et al., 1986; Frith et al., 1994; Garnham and Ruffman, 2001; Gopnik and Astington, 1988; Happé, 1994; Hogrefe et al., 1986; Jolliffe and Baron-Cohen, 1999; Klin, 2000; Muris et al., 1999; O’Neill, 1996, 2005; Perner et al., 1987; Ruffman et al., 2001; Schneider et al., 2012a; Tager-Flusberg, 1992; Wellman and Estes,
1986; White et al., 2009; Wimmer and Perner, 1983; Zaitchik, 1991); several of these studies were reviewed in Chapter 1. For the purposes of this study, it was necessary to identify methods of ToM assessment that were short, applicable to neurotypical adults, and readily scored. Three methods that meet these criteria are the Autism-spectrum Quotient, the Reading the Mind in the Eyes test, and the Strange Stories task.

The Autism-spectrum Quotient (AQ; Baron-Cohen et al., 2001b) is a common method for assessing autistic traits in adults of normal intelligence. The task takes the form of a short self-reported questionnaire, providing statements such as “I am good at social chit-chat” to which the test-taker responds on a four-point agree/disagree scale. A higher score on the AQ indicates more autistic traits. In the original scoring system outlined by Baron-Cohen et al. (2001b), each response is either scored as 1 point, for an ‘autistic’ response, or 0 points, for a ‘non-autistic’ response. In other words, the distinction between ‘strongly agree’ and ‘agree’ is erased; the strength of the response does not matter, only the direction. More recent work, however, uses a Likert system, assigning 1–4 points for each response, with 4 for the most extreme ‘autistic’ response and 1 for the most extreme ‘non-autistic’ response. This scoring system was adopted by Austin (2005) and Hoekstra et al. (2008, 2007), and subsequent versions of the AQ generally instruct Likert-style response coding (such as the AQ-Child, designed for use with children; see Auyeung et al., 2008). Due to its ease of administration and scoring and high test–retest reliability, the AQ has seen wide use in the literature since its introduction (but cf. Bishop and Seltzer, 2012; Murray et al., 2014). The AQ was designed to assess five specific subscales, based on classic symptoms of autism: communication, social skills, attention switching, imagination, and attention to detail, but subsequent psychometric research (Austin, 2005; Broadbent et al., 2013; Eriksson, 2013; Hoekstra et al., 2008, 2011; Hurst et al., 2007; Kloosterman et al., 2011; Lau et al., 2013; Stewart and Austin, 2009, 2010) has suggested alternative subscale structures based on
factor analyses of responses. Regardless of the fine structure of the subscales, however, the full AQ score has been demonstrated to be reliable (e.g. Austin, 2005), compare favorably with other measures of autistic traits (Armstrong and Iarocci, 2013), and not be easily explainable in terms of personality traits (Wakabayashi et al., 2006), suggesting that the AQ does indeed measure cognitive style rather than simply idiosyncratic personal preferences.

The AQ was designed to be used to measure autistic traits, not theory of mind. However, to the extent that the central cognitive deficit of autism is a lack of ToM, the AQ can be used as a measure of ToM. Indeed, in Senju et al.’s (2009) study of anticipatory eye movements in people with Asperger syndrome during a false belief task, the groups of participants were matched as much as possible in terms of IQ and verbal ability. Participants also underwent a battery of ToM tests and completed the AQ questionnaire. Among these measures, the only significant difference between the groups was in AQ score—the Asperger syndrome group had much higher AQ scores than the NT group—i.e. no differences between groups in the ToM tasks were observed. Nevertheless, the experimenters still reported finding significant results in terms of eye movements, revealing that the Asperger syndrome group had a ToM deficit relative to the NT group. Since AQ was the only metric differentiating the groups (aside from the clinical diagnoses), then, the AQ scores serve as a measure of ToM, capturing that which the other tests were unable to adequately assess.

In the Reading the Mind in the Eyes (RMITE) test, the participant is presented with a photograph of the eye region of the face (see Figure 2.2 for an example), and asked which of four adjectives best describes what the person in the photograph is feeling (Baron-Cohen et al., 1997a, 2001a, 1997b). This test assesses the ability to infer complex emotional states from simple pictures, and thus serves as a measure of theory of mind. It has been used in both autism studies and more general psychological
investigations of ToM (e.g. Sylwester et al., 2012), and has been shown to be a reliable and stable measure of ToM in non-clinical adult populations (Fernández-Abascal et al., 2013; Vellante et al., 2013).

Finally, in the Strange Stories task, the participant reads (or is read) short, paragraph-length stories and answers a question about the motivations of and relationships between the story’s characters (Happé, 1994; Jolliffe and Baron-Cohen, 1999; White et al., 2009). For example, participants are asked why a particular character told a lie or did some unusual action. Their responses are scored based on whether or not the answer was correct and whether or not the answer made reference to the mental state of others. This test is able to distinguish ToM ability on the basis of how often participants make reference to the mental states of others versus reference to the physical world.
2.1.5 Effects of autistic traits on linguistic variables

The effects of individual differences in autistic traits on language use has received increased attention within linguistics recently. Many of these studies have attempted to correlate autistic traits (measured via AQ scores) with particular linguistic behaviors (Bishop, 2012; Jun and Bishop, 2014, 2015; Stewart and Ota, 2008; Yu, 2010, 2013; Yu et al., 2013, 2011), while others have examined the linguistic behaviors of individuals with autism spectrum conditions (Clopper et al., 2012, 2013; Mielke et al., 2013). We will discuss one study, by Yu (2010), in some depth, as it is replicated in the present chapter.

Yu’s (2010) study examined perceptual accommodation to coarticulation. The fricative /ʃ/ has a lower centroid frequency than the fricative /s/, and in natural speech, these fricatives have a lower frequency when produced before /u/ than /a/ (due to lip protrusion causing a larger front cavity). Therefore, /s/ in a /u/ context sounds more /ʃ/-like than it does in a /a/ context; likewise, /ʃ/ in a /a/ context sounds more /s/-like than it does in a /u/ context. Listeners are aware of this influence of coarticulation on acoustics, and adjust their category boundaries accordingly (Mitterer, 2006); that is, the perception of the same acoustic token can flip between ‘s’ and ‘sh’ depending on the vocalic context it is presented in. In Yu’s (2010) experiment, participants were presented with a recording of a CV syllable and asked to classify the consonant as either ‘s’ or ‘sh’. The stimuli consisted of a seven-step continuum of /s/ to /ʃ/ spliced onto tokens of /a/ and /u/. Yu (2010) found that, for the female participants, the extent to which the consonant percept is influenced by the vowel is correlated with the participant’s AQ score. That is, women with a higher AQ score (i.e. more autistic traits) were more likely to be influenced by vowel identity in their classification of the fricative than were the women with a lower AQ score.
Interpreting these results is difficult. The women with the low AQ scores attended more to the within-category variation in the fricative spectra than the women with the high AQ scores, who made use of the vowel information instead. In light of the literature on the ability of people with autism to attend to context, these findings are counterintuitive. There is ample evidence that people with autism show great attention to physical detail, while often missing contextual or global cues (Happé and Frith, 2006). For instance, compared to neurotypical controls, people with autism are less able to recognize global properties of speech such as emotional content (Kleinman et al., 2001; Stewart et al., 2011) or dialect (Clopper et al., 2012), and rely less on context to disambiguate homographs (Happé, 1997). Further, among people with autism there is a greater incidence of absolute pitch perception (Heaton et al., 1998), and an observed bias toward acoustic attention over linguistic attention (Järvinen-Pasley et al., 2008). Given these facts, it seems reasonable to expect that listeners with more autistic traits are more likely to attend to small, local details such as the absolute frequency of fricative noise, and less likely to attend to larger, global details such as the identity of the following vowel, especially in the context of a task where the stated objective is to judge the identity of the fricative consonant.

However, the results reported by Yu (2010) show the opposite pattern. Female listeners with more autistic traits were more likely to be influenced by the vocalic context than female listeners with fewer autistic traits. Yu (2010) framed the result as being due to the fact that people with more autistic traits have (relatively) impoverished social representations (Klin, 2000), and that fewer resources attending to (high-level) social cues means more resources available for (low-level) perceptual compensation. Further, individuals with more autistic traits exhibit an enhanced ability to ‘systemize’, to create associations between objects and rules (Baron-Cohen et al., 2002). Unfortunately, this post hoc explanation only serves to illustrate the fact that, since autism
is an extremely heterogeneous syndrome, it may be possible to predict any number of seemingly contradictory results based on trends reported in the literature.\(^2\)

The male listeners in Yu’s (2010) study exhibited perceptual accommodation—i.e., attended to vocalic information in their phoneme classification—regardless of their AQ score. In this regard, their performance was similar to that of the high AQ female listeners. From this perspective, the exceptional group whose performance should be accounted for is the group of low AQ women. One possible sociophonetic interpretation of this result is that the women with the low AQ scores, being necessarily skilled in social interaction and cognizant of social meaning, attend to the fricative because it is the fricative that bears socioindexical meaning, not the vowel (Babel, 2012; Strand, 1999).\(^3\) An alternative account appeals to findings from speech perception: McGuire (2007) demonstrated that English listeners, when attempting to distinguish the Polish sibilants /ʃ/ and /c/, mainly relied on formant transition information rather than the more reliable spectral content of the fricative noise. Simply put, distinguishing fricative noise is ‘more difficult’ than vowel formants, especially for listeners unused to making such a distinction (see also Nittrouer, 2002; Nittrouer and Miller, 1997). Under this account, the women with the higher AQ scores and the men have developmentally ‘more mature’ perceptual systems, and are able to quickly parse the vowel and fricative information. This reasoning predicts that people with higher AQ scores are more able to adjust to speech in adverse conditions (but cf. Alcántara et al., 2004).

Given that individual variation in autistic traits (and ex hypothesi, ToM) has been shown to influence both language production and perception, it is possible that ToM correlates with the extent of predictability-based phonetic reduction, as predicted

\(^2\)Indeed, phonological categories and their relationship to social categories are also very complex phenomena, further complexifying the possible predictions.

\(^3\)However, this interpretation is difficult to align with our understanding of social meaning carried by vowels, such as in /u/- and /a/-fronting.
by a strong version of the listener-oriented account. The experiments detailed below investigated this hypothesis—specifically, whether or not there is a correlation between individual differences in ToM and the five kinds of phonetic reduction discussed above.

### 2.2 Experiment 1a: Semantic predictability-based reduction and second mention reduction

#### 2.2.1 Introduction

The first experiment investigated whether or not there is a correlation between individual differences in ToM and semantic predictability-based phonetic reduction or second mention reduction.

#### 2.2.2 Method

**2.2.2.1 Participants**

Twenty-one eligible participants completed the task for partial course credit. Eligibility was restricted to native monolingual speakers of American English with no history of speech, language, or hearing disorders. None of the participants reported any history of autism-spectrum conditions.

**2.2.2.2 Procedure**

Participants completed three linguistic tasks and three ToM assessment tasks, in that order, all administered via a computer in a double-walled sound attenuated booth. In the first task, the participant was instructed to read sentences aloud. The sentences
were 41 matched high- and low-predictability (HP and LP) sentence pairs, drawn from the SPIN sentences (Kalikow et al., 1977), and varied in the semantic predictability of the final (target) word. An example pair is shown in (3) and (4); the full list of sentences used is provided in Appendix A. The sentences were presented randomly for each participant, blocked by predictability condition. Block order was counterbalanced between participants. Each sentence was presented individually on a computer screen for 3.5 seconds, before advancing to the next sentence. This feature therefore made the task a speeded one, although participants reportedly found the pace quite comfortable. This task was intended to induce semantic predictability based reduction.

(3) For your birthday I baked a cake.

(4) Tom wants to know about the cake.

In the next task, the participant read five paragraphs adapted slightly from those developed by Baker and Bradlow (2009). These paragraphs featured repetitions of target items in different contexts and were intended to elicit second mention reduction. These paragraphs are provided in full in Appendix B. The paragraphs had a total of 40 target words, each said twice, for a total of 80 target word tokens. Each paragraph was presented on screen, and the trial advancement was self-paced.

The third task was a replication of Yu’s (2010) phoneme identification task, detailed above. Fricative-vowel tokens were presented over professional monitoring headphones and participants classified the consonant as either ‘s’ or ‘sh’ via a button box. Stimulus construction followed the procedure outlined by Yu (2010). Briefly, a seven-step /s-ʃ/ continuum was created by mixing weighted averages of the waveforms of productions of /sæ/ and /ʃæ/ from a female native speaker of American English. These fricatives were then cross-spliced with the vocalic portions of /æ/ and /u/ taken from productions of /dæ/ and /du/ produced by a male and a female native speaker.
of American English. The stimuli therefore involved 7 steps, 2 vowels, and 2 talker genders, for a total of 28 unique tokens. Figure 2.3 depicts centroid values at each continuum step, taken over the entire fricative portion of each stimulus.

This task was included as a method of assessing the usefulness of each of the ToM measures. Assuming that Yu’s (2010) result can be replicated, it may be possible to use the RMITE score or the Strange Stories score to stand in for AQ as a measurement of ToM. If so, then those ToM scores can be considered functionally equivalent to AQ, at least for the purposes of perceptual accommodation. This equivalence is methodologically useful in establishing which of these three measures of ToM are informative of variation in the participant population.
Following the linguistic tasks, the participants completed the ToM tasks. The first of these was the RMITE task discussed above. It has been demonstrated that people with autism spectrum conditions and other ToM deficits perform significantly worse than typically developing controls on this task (Baron-Cohen et al., 1997a, 2001a). The next task was the Autism-spectrum quotient (AQ) questionnaire, and the final task was the Strange Stories task.

Both the RMITE and AQ tasks were presented as they were published. However, the Strange Stories task involves verbal descriptions of scenarios with a distinctively British character (both in terms of syntax and lexis), which could distract or confuse an American participant pool. Therefore, each of the stories and questions were rephrased into a format that American English speakers found more natural. This rephrasing was performed in consultation with a team of six native speakers of American English and one native speaker of British English. Full story and question sets are provided in Appendix C.

2.2.2.3 Acoustic measurements

Measures of word duration, vowel duration, and vowel midpoint F1 and F2 were taken of the target words of both production tasks. The duration measures were taken in milliseconds, and the formant measures were taken in Hz using standard LPC analysis with peak picking and then converted to ERB. The formant values were used to calculate vowel dispersion which was defined for each token as the Euclidean distance from that token to the talker-specific vowel center. The vowel center was defined as the grand mean of by-vowel means for that talker.

---

4ERB, or equivalent rectangular bandwidth, is a psychoacoustically-motivated scale, similar to the Bark scale.
In the sentence-reading task, each target word was at the end of the sentence, and due to the elicitation procedure, also necessarily phrase-final. In the paragraph-reading task, however, many of the words were sentence-medial, and may or may not have been followed by a prosodic break. The presence or absence of a pause after every target word was therefore also coded, to partially control for effects of phrase-final lengthening.

2.2.2.4 Analysis

Speech tokens were excluded if they were disfluent or restarted during speech. For the sentence-reading task, out of a total of 1,722 target tokens (21 participants × 82 target tokens), 72 tokens were excluded; for the paragraph-reading task, from a total of 1,680 target tokens (21 participants × 80 target tokens), 63 tokens were excluded. Therefore, the analysis of the sentence-reading includes 1,650 tokens, while the paragraph-reading analysis includes 1,617 tokens.

A linear mixed effects regression model was constructed for each production task for each of the acoustic variables—word duration, vowel duration, and vowel dispersion. Fixed effects were predictability condition or discourse mention (high predictability versus low predictability, with high as reference level; or first mention versus second mention, with first as reference level), AQ score, RMITE score, and 2-way interactions between predictability/mention and AQ, and predictability/mention and RMITE. All continuous variables were centered around the mean before being entered into the model. Random intercepts for talker and word identity were included, with random slopes for predictability condition for both talker and word. In all models, $p$-values were calculated by treating the absolute $t$-statistic as if it came from a $t$-distribution with degrees of freedom equal to the number of data points minus the number of model parameters (Baayen, 2008).
Because the phoneme classification task was a close replication of Yu's (2010) study, a logistic mixed effect regression model was constructed according to the specifications described by Yu (2010). The model contained six fixed effects: trial, continuum step (1–7), talker gender (male or female), vowel (/a/ or /u/), participant gender (male or female), and total log AQ.\(^5\) Additionally, six two-way interactions of voice gender × step, vowel × step, voice gender × vowel, participant gender × vowel, log AQ × vowel, and participant gender × vowel; and a single three-way interaction of participant gender × log AQ × vowel. Participant identity was entered as a random effect, with a random slope of trial.\(^6\) All continuous variables were centered at zero, and all discrete variables were coded with sum contrasts. Additionally, to validate the use of the RMITE and Strange Stories tasks, two additional models were constructed, substituting the RMITE or Strange Stories score in place of AQ score. Experimenter error led to the phoneme classification task not being carried out for one participant. Therefore, there were 2,240 responses (20 participants × 112 trials) available for analysis.

2.2.2.5 Predictions

If the strong interpretation of the listener-oriented account is correct, then it is expected that participants with better theory of mind abilities ought to produce larger and more consistent differences between high- and low-predictability items and between first and second mentions than participants with poorer theory of mind. This prediction is due to the fact that, according to this strong interpretation, a poor theory of mind entails

---

\(^5\)Yu (2010) log-transformed the AQ scores. This transformation is not conventional in autism research, but it is repeated here for the sake of replication. All subsequent analyses of AQ in this dissertation use the untransformed AQ score.

\(^6\)This differs from the original analysis in one respect: in the original, trials were split into arbitrary blocks, and block was entered as a fixed effect. In the present study, the trials were not blocked, and thus it was not possible to include this factor.
poorer mental interlocutor models, and thus less ability to predict which linguistic items need to be enhanced.

Therefore, if the results indicate that degree of reduction (i.e. durational difference between high predictability and low predictability conditions, or first and second mentions) correlates positively with ToM, then these results can be interpreted in favor of the listener-oriented account. If there is no such correlation, then the strong interpretation of the listener-oriented account may not hold.

2.2.3 Results

2.2.3.1 Individual difference scores

AQ scores ranged from 90 to 128, with a mean of 111.48 ($SD = 11.02$), RMITE scores ranged from 15 to 33, with a mean of 26.92 ($SD = 4.77$), and the raw scores on the Strange Stories task ranged from 16 to 53, with a mean of 42.62 ($SD = 8.23$). These AQ and RMITE figures are not far from the expected population means established in prior work (Austin, 2005; Baron-Cohen et al., 2001a). There are no published expected population means for the Strange Stories score for comparison.

In the Strange Stories task, as well as questions about the mental states of individuals in the story, there were also control questions which were essentially text comprehension questions. The inclusion of these questions in the task makes it possible to determine if a participant actually has a poor ToM (indicated by low scores on the mental state questions but not on the comprehension questions) or is just a poor test-taker, or not paying much attention during the experiment (indicated by low scores on the comprehension questions as well). Therefore, a Strange Stories score difference (SSSD) was calculated by subtracting the mental question scores from the comprehension question scores. Thus, a score near zero indicates that the participant performed
Table 2.2: Correlation matrix of the individual difference scores in experiment 1a, also including p-values for the significance of each pairwise comparison.

<table>
<thead>
<tr>
<th></th>
<th>AQ score</th>
<th>RMITE</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r (p)</td>
<td>r (p)</td>
</tr>
<tr>
<td>RMITE score</td>
<td>-.154 (.505)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strange Stories</td>
<td>-.383 (.087)</td>
<td>.673 (.001)</td>
<td></td>
</tr>
<tr>
<td>SSSD</td>
<td>.129 (.578)</td>
<td>.163 (.480)</td>
<td>-.24 (.295)</td>
</tr>
</tbody>
</table>

As well on the mental state questions as on the comprehension questions, while a large positive score indicates that the participant performed poorly on the mental state questions, relative to the comprehension questions. The SSSD scores ranged from -1 to 7, with a mean of 3.10 (SD = 2.23). Table 2.2 depicts a correlation matrix between all four of these individual difference scores, showing both Pearson’s r and the p-value. As can be seen, the only scores that were significantly correlated were the RMITE and plain Strange Stories scores.\(^7\)

2.2.3.2 Phoneme classification results

The output of the model specified by Yu (2010) is shown in Table 2.3. The expected effects of and interactions between continuum step, talker gender, and vowel were observed; these effects are visualized in Figure 2.4. Tokens with more [ʃ]-like acoustics were given more ‘sh’ responses; tokens from the male talker received fewer ‘sh’ responses than those from the female talker; and tokens with /u/ received fewer ‘sh’ responses than those with /ɑ/. Unlike Yu’s results, however, the interactions between AQ and vowel, and between AQ, listener gender, and stepsize were not significant.

\(^7\)This correlation is significant even after applying a Bonferroni correction: \(\alpha = .05/6 \approx .008 > .001\).
Table 2.3: Coefficients and significance values for logistic model predicting response in the phoneme classification task. LG = “Listener gender”; TG = “Talker gender”.

The results of the model substituting RMITE score for log AQ was qualitatively similar to the original AQ model. Crucially, however, in this model the vowel × RMITE score interaction was significant ($\beta = 0.095$, $z = 3.565$, $p < .001$), as was the vowel × RMITE × listener gender interaction ($\beta = -0.116$, $z = -2.179$, $p = .029$). These interactions are visualized in Figure 2.5; the responses of low-RMITE female listeners were more sensitive to vowel context than the responses of other listeners. This result mirrors Yu’s (2010) finding of heightened sensitivity in high-AQ listeners. Recall here that a higher AQ score indicates a greater prevalence of autistic traits (and, by hypothesis, a poorer ToM), while a lower RMITE score indicates a poorer ToM; that is, in terms of ToM, these results are consistent with Yu’s (2010) reported results. Neither of these critical interactions reached significance in the Strange Stories score model. Taken together, these results suggest that the RMITE scores and the AQ scores are measuring
Figure 2.4: Interactions between vowel context, talker gender, and continuum step in responses to the phoneme classification task.

similar dimensions of individual differences, but that the Strange Stories score is either not accurately measuring individual variation in ToM, or that it is measuring an aspect of ToM not involved in perceptual accommodation to coarticulation.

2.2.3.3 Semantic predictability-based reduction results

The output of the regression analysis predicting word duration in the sentence-reading task is summarized in Table 2.4. A simple effect of predictability condition was observed, such that unpredictable words ($M = 429$ms) were longer than predictable words ($M = 410$ms). This result is consistent with prior research (Lieberman, 1963; Turnbull and Clopper, 2013), and demonstrates that phonetic reduction occurred. No simple effects of the individual difference scores (AQ and RMITE) were observed, but
significant interactions were observed. AQ score interacted with predictability, such that participants with higher AQ scores had a larger difference in word duration between the predictable and unpredictable contexts than participants with lower AQ scores. That is, these participants had a larger extent of phonetic reduction than low-AQ participants. A similar interaction was observed with the RMITE scores, whereby participants with higher RMITE scores had shorter words in the unpredictable condition than participants with lower RMITE scores. After recalling that higher scores in the AQ are interpreted the same as lower scores on the RMITE, a clear pattern emerges: participants with less ToM ability have larger or more extensive reduction in word duration between the low and high predictability conditions. The interaction between predictability and AQ is visualized in Figure 2.6, and that of predictability and RMITE

Figure 2.5: Interaction between vowel context and participants’ RMITE score in responses to the phoneme classification task, split by listener gender.
is visualized in Figure 2.7. Note that the x-axis of Figure 2.7 is inverted so that, in all the graphs, participants with poorer ToM skills cluster toward the right, while those with better ToM skills cluster toward the left.

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-8.207</td>
<td>-0.447</td>
<td>.655</td>
</tr>
<tr>
<td>Pred.: Low</td>
<td>19.874</td>
<td>3.885</td>
<td>&lt;.001  **</td>
</tr>
<tr>
<td>AQ score</td>
<td>0.723</td>
<td>0.301</td>
<td>.763</td>
</tr>
<tr>
<td>RMITE score</td>
<td>1.976</td>
<td>0.906</td>
<td>.365</td>
</tr>
<tr>
<td>Pred. × AQ</td>
<td>2.759</td>
<td>2.425</td>
<td>.015   *</td>
</tr>
<tr>
<td>Pred. × RMITE</td>
<td>-3.052</td>
<td>-2.943</td>
<td>.003 **</td>
</tr>
</tbody>
</table>

Table 2.4: Model output for word duration model in the sentence production task.

The output of the regression analysis predicting vowel duration in the sentence-reading task is summarized in Table 2.5. A simple effect of predictability condition was observed, such that vowels in unpredictable words ($M = 188$ms) were longer than vowels in predictable words ($M = 181$ms). No other significant simple effects or interactions were observed. However, note that the interactions between AQ and predictability and between RMITE and predictability are in the same direction as those observed in the word duration model. That is, higher AQ and lower RMITE scores both correlated with a greater extent of phonetic reduction (duration differences between high and low predictability conditions). However, given the lack of a statistically significant result, this trend cannot be reliably interpreted.

Table 2.6 summarizes the output of the regression analysis predicting vowel dispersion in the sentence-reading task. No significant effects were observed, suggesting that the observed semantic predictability-based reduction was mostly temporal.
Figure 2.6: Reduction in word duration from low to high predictability conditions as a function of talker AQ score. A positive duration difference means the LP production was longer than the HP production. This and all subsequent graphs in this section depict subject means and standard errors with linear trend and confidence interval overlaid.

Taken together, these results suggest that individuals with less theory of mind (higher AQ scores, lower eye reading scores) in fact exhibit larger predictability effects than individuals with more theory of mind. This conclusion is unexpected: the listener-oriented account predicts that poorer theory of mind should lead to less listener adaptation, not more. However, interpreting these results in light of Yu’s (2010) results may provide a clearer picture. Recall that Yu (2010) found that higher AQ scores led to more perceptual adaptation. This result was replicated in the present data, for both AQ and RMITE scores. Yu (2010) interpreted this result as a reflection of the increased pattern recognition and systematic association abilities of people with higher AQ, which leads to greater perceptual adaptation. Similarly, those with a higher AQ are better
able to generalize longer durations to unpredictable novel sentences. This explanation suggests that our current understanding of the AQ and its relationship to ToM and listener modeling is inadequate and deserves further investigation.
Table 2.6: Model output for vowel dispersion model in the sentence production task.

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.003</td>
<td>0.013</td>
<td>.990</td>
</tr>
<tr>
<td>Pred.: LP</td>
<td>$-0.018$</td>
<td>$-0.484$</td>
<td>.628</td>
</tr>
<tr>
<td>AQ</td>
<td>0.015</td>
<td>0.762</td>
<td>.446</td>
</tr>
<tr>
<td>RMITE</td>
<td>0.024</td>
<td>1.367</td>
<td>.172</td>
</tr>
<tr>
<td>Pred. $\times$ AQ</td>
<td>0.001</td>
<td>0.087</td>
<td>.931</td>
</tr>
<tr>
<td>Pred. $\times$ RMITE</td>
<td>$-0.012$</td>
<td>$-1.590$</td>
<td>.112</td>
</tr>
</tbody>
</table>

2.2.3.4 Second mention reduction results

The model outputs from the word duration, vowel duration, and vowel dispersion models from the paragraph-reading task are shown in Tables 2.7, 2.8, and 2.9, respectively. As can be seen, significant effects of mention were observed for all three acoustic variables. Words were significantly shorter on their second mention ($M = 313$ms) than on their first mention ($M = 337$ms), vowels were similarly shorter and less disperse on the second mention ($M = 126$ms, $M = 2.778$ERB) relative to the first mention ($M = 115$ms, $M = 2.884$ERB). These effects are consistent with prior literature on second mention reduction (Baker and Bradlow, 2009; Burdin and Clopper, 2015). Additionally, consistent with the effects of phrase-final lengthening (Burdin and Clopper, 2015; Turk and Shattuck-Hufnagel, 2000, 2007), words preceding a pause were significantly longer ($M = 378$ms) than words not preceding a pause ($M = 315$ms), and vowels in words preceding a pause were significantly longer ($M = 143$ms) than vowels in words not preceding a pause ($M = 116$ms). No interactions with any of the ToM variables were observed in any model.
<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.994</td>
<td>0.223</td>
<td>.824</td>
</tr>
<tr>
<td>Word is pre-pausal</td>
<td>55.147</td>
<td>10.236</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Mention: Second</td>
<td>-24.401</td>
<td>-4.601</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>AQ</td>
<td>0.317</td>
<td>0.427</td>
<td>.669</td>
</tr>
<tr>
<td>RMITE</td>
<td>-0.710</td>
<td>-0.323</td>
<td>.747</td>
</tr>
<tr>
<td>Strange Stories</td>
<td>-0.265</td>
<td>-0.184</td>
<td>.854</td>
</tr>
<tr>
<td>SSSD</td>
<td>-4.666</td>
<td>-1.255</td>
<td>.210</td>
</tr>
<tr>
<td>Mention × AQ</td>
<td>-0.261</td>
<td>-0.800</td>
<td>.424</td>
</tr>
<tr>
<td>Mention × RMITE</td>
<td>0.116</td>
<td>0.119</td>
<td>.905</td>
</tr>
<tr>
<td>Mention × Strange Stories</td>
<td>-0.303</td>
<td>-0.453</td>
<td>.651</td>
</tr>
<tr>
<td>Mention × SSSD</td>
<td>-1.151</td>
<td>-0.687</td>
<td>.492</td>
</tr>
</tbody>
</table>

Table 2.7: Output of word duration model in the paragraph-reading task.

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.584</td>
<td>0.656</td>
<td>.512</td>
</tr>
<tr>
<td>Word is pre-pausal</td>
<td>13.210</td>
<td>5.247</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Mention: Second</td>
<td>-11.641</td>
<td>-4.218</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>AQ</td>
<td>-0.016</td>
<td>-0.045</td>
<td>.964</td>
</tr>
<tr>
<td>RMITE</td>
<td>-0.206</td>
<td>-0.195</td>
<td>.846</td>
</tr>
<tr>
<td>Strange Stories</td>
<td>-0.481</td>
<td>-0.693</td>
<td>.488</td>
</tr>
<tr>
<td>SSSD</td>
<td>-1.807</td>
<td>-1.011</td>
<td>.312</td>
</tr>
<tr>
<td>Mention × AQ</td>
<td>-0.014</td>
<td>-0.081</td>
<td>.935</td>
</tr>
<tr>
<td>Mention × RMITE</td>
<td>0.369</td>
<td>0.730</td>
<td>.466</td>
</tr>
<tr>
<td>Mention × Strange Stories</td>
<td>-0.298</td>
<td>-0.865</td>
<td>.387</td>
</tr>
<tr>
<td>Mention × SSSD</td>
<td>-1.166</td>
<td>-1.343</td>
<td>.179</td>
</tr>
</tbody>
</table>

Table 2.8: Output vowel duration model in the paragraph-reading task.
<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.053</td>
<td>0.265</td>
<td>.791</td>
</tr>
<tr>
<td>Word is pre-pausal</td>
<td>0.049</td>
<td>1.057</td>
<td>.291</td>
</tr>
<tr>
<td>Mention: Second</td>
<td>-0.115</td>
<td>-4.681</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>AQ</td>
<td>0.003</td>
<td>0.398</td>
<td>.691</td>
</tr>
<tr>
<td>RMITE</td>
<td>0.009</td>
<td>0.412</td>
<td>.681</td>
</tr>
<tr>
<td>Strange Stories</td>
<td>0.006</td>
<td>0.418</td>
<td>.676</td>
</tr>
<tr>
<td>SSSD</td>
<td>-0.009</td>
<td>-0.241</td>
<td>.809</td>
</tr>
<tr>
<td>Mention $\times$ AQ</td>
<td>-0.000</td>
<td>-0.048</td>
<td>.961</td>
</tr>
<tr>
<td>Mention $\times$ RMITE</td>
<td>-0.007</td>
<td>-1.015</td>
<td>.310</td>
</tr>
<tr>
<td>Mention $\times$ Strange Stories</td>
<td>-0.005</td>
<td>-1.045</td>
<td>.296</td>
</tr>
<tr>
<td>Mention $\times$ SSSD</td>
<td>-0.014</td>
<td>-1.127</td>
<td>.260</td>
</tr>
</tbody>
</table>

Table 2.9: Output of vowel dispersion model in the paragraph-reading task.
2.2.4 Discussion

2.2.4.1 Equivalence of individual difference scores

Both the RMITE and AQ scores were observed to be useful in explaining variance in the phoneme classification task, but the Strange Stories task was not. Part of this discrepancy may be related to task demands. Both the RMITE and AQ tasks require the participant to choose one of four possible responses to a limited task. In the RMITE task, participants must ascribe an emotion adjective to the picture; in the AQ task, participants must rate how much they agree with a short sentence. The Strange Stories task, on the other hand, requires greater levels of heightened attention. Participants must read a story, multiple sentences in length, and attend to several details. Their response to the question is entirely free-form: they can respond with a single word or a whole essay, but the range of possible responses is effectively infinite. The question asked for each story is quite different with each trial. Given these constraints, it is possible that some participants receive low scores simply due to putting a low level of effort into their answers, or because they were unsure of the necessary level of detail required in the responses. Scoring the answers is also a somewhat subjective affair, despite the clear rubric, and is an additional source of variance in these scores. Finally, the task was designed for use primarily with children, while the AQ and RMITE task were both designed for adults. The often trivially-easy questions in some filler items of the Strange Stories task may have perturbed the adult undergraduate participants. Due to these concerns, the significant amount of experiment time taken up in administering the task, and the significant amount of researcher time taken in scoring the responses, the Strange Stories task will not be used in any further experiments.
The results of the replication of Yu’s (2010) perceptual accommodation task suggest that AQ and RMITE scores can be used somewhat interchangeably in the modeling of individual differences in accommodation to coarticulation. This fact does not immediately translate to the conclusion that they measure the same construct, or even that they will be interchangeable in their effects on a different linguistic task. Nevertheless, the results do hint at some degree of equivalence between these scores.

### 2.2.4.2 Phonetic reduction and theory of mind

A strong version of the listener-oriented theory of predictability-based phonetic reduction requires talkers to possess a well-developed ToM in order to implement reduction (and the lack of reduction) in an appropriate manner. Under this interpretation, individuals with less adept ToM skills should produce phonetic reduction less consistently than individuals with more adept ToM skills. This experiment tested two domains of phonetic reduction—semantic predictability-based reduction and second mention reduction—and revealed interactions between ToM skill and semantic predictability-based reduction, but not with second mention reduction. However, the interaction was not in the direction predicted by listener-oriented theories: talkers with poorer ToM in fact produced larger acoustic differences between the predictability conditions than talkers with better ToM, implying that ToM abilities somehow hinder the effective use of reduction in semantically predictable contexts.

The other main result of the experiment was that second mention reduction was not observed to interact with ToM skills in any way. All participants, regardless of ToM assessment scores, produced second mention reduction to roughly the same degree. Additionally, while the semantically predictable words underwent only temporal reduction, the second-mentioned words underwent both temporal and spectral reduction. Taken together, these results from acoustics and individual differences suggest
that second mention reduction and semantic predictability-based reduction arise from different cognitive mechanisms, despite their apparent reliance on discourse-sensitive contextual information.

The unexpected effect of AQ on semantic predictability-based reduction defies both listener-based and talker-based understandings of phonetic reduction. One potential recourse is to deny that the effect is actually due to AQ score, but rather attribute it to some underlying variable. One candidate variable for this explanation is general intelligence, conventionally quantified by IQ (intelligence quotient). Bishop and Seltzer (2012) presented evidence that IQ weakly correlates with AQ in individuals with autism spectrum disorders (but cf. Rajkumar et al., 2008), and Hoekstra et al.’s (2010) longitudinal twin study suggested that a small genetic link between autism and intelligence exists. More generally, studies of language in individuals diagnosed with autism tend to use IQ-matched neurotypical individuals as controls, although this practice is becoming regarded as based on somewhat naïve assumptions about the relationship between autism and intelligence (Tager-Flusberg, 2004). Therefore, it is possible that the high-AQ participants in this experiment also happened to have high IQ scores, and therefore, with greater analytic skill, are quickly able to determine the predictability of each sentence as they read it. The lower AQ participants, with lower IQ, are not able to process the sentences as quickly, and thus produce all sentences the same. Recall here that the reading task was speeded—participants only had 3.5 seconds to read each sentence. An alternative or supplementary reason could also be that participants with lower IQ are simply poor readers, relative to higher IQ participants. Such an explanation requires obtaining a measure of general intelligence. Experiment 1b therefore replicates the sentence-reading task, and includes a task designed to estimate IQ.
2.2.4.3 Consciousness and control

A related question that arises from these results is the extent to which any of these observed phonetic reductions are controlled by the talker, or if they are automatic, arising from unconscious processes. The study of consciousness and control of action is complex and relatively interdisciplinary (see Umiltà, 2007, for review), although the brief review in this paragraph focuses on social psychology (rather than, say, metacognition or limb control) due to its clear connections to language use. Under the classical two-process model of automatic versus controlled processing (Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977), automatic processes are effortless, unconstrained by capacity limitations such as working memory, and react to stimuli without control or action. Controlled (or conscious) processes, however, are effortful and constrained by capacity limitations, but are relatively flexible. The accuracy of controlled processes can be improved through practice, and their progress can be interrupted by automatic processes. Automatic processes cannot be interrupted nor improved through practice. Automatic processes are said to occur outside of conscious awareness, be unintentional, be efficient in their use of resources, and be uncontrollable, while controlled processes are the opposite. In the decades following Schneider and Shiffrin’s proposal, research in social cognition converged on the consensus that this dichotomy was essentially a continuum, and that many processes can be described as having properties of both an automatic and a controlled process (Bargh, 1984, 1994, 2007; Bargh and Chartrand, 1999). More recent research has also called into question the extent to which the mental representation of desired outcomes, volitional will, and consciousness are separable units (Dijksterhuis and Aarts, 2010; Moors and De Houwer, 2006), generally concluding that many processes can proceed both with and without conscious attention.
In terms of speech production, there is a similar debate in the literature about the extent to which particular acoustic cues can be controlled versus what can be ascribed to ‘automatic’ effects resulting from the physiology of the speech apparatus (see Solé, 2007, for review). A related area of study is that of ‘clear speech’, speech produced under adverse conditions such as heavy background noise or directed to a hearing-impaired listener (see Smiljanić and Bradlow, 2009, for review). The relationship between such explicit speech control and processes assumed to be ‘automatic’, such as phonetic reduction, is relatively understudied. There is evidence that performance in a trained motor skill can be negatively affected by focusing attention on components of the action (see Beilock and Carr, 2001, and the references therein); it is thought that the conscious attention overrides the more efficiently-organized automatic processes involved. From this perspective, it might be expected that many ‘automatic’ processes fail to occur in clear speech, where the talker is effortfully trying to make their speech more clear. However, in one of the few investigations of these questions, Baker and Bradlow (2009) observed both frequency effects—phonetic reduction on more frequent words relative to less frequent words—and second mention reduction in clear speech, suggesting that the mechanisms of clear speech do not necessarily ‘override’ these processes. The relationship between clear speech and semantic predictability-based reduction, however, remains unstudied.

Additionally, the literature on clear speech has established that there is considerable variation between talkers in their strategies and effectiveness in implementing clear speech (Ferguson, 2004, 2012). The extent to which these strategies can be related to measurable patterns of individual differences in cognition remains understudied. A possible hypothesis relating clear speech to theory of mind is that individuals with better ToM produce greater or more effective enhancements in clear speech, due to their ability to put themselves in the shoes of their interlocutor. Individuals
with poorer ToM, on the other hand, would implement clear speech in a less effective manner. Experiment 1b considers this possibility by examining interactions between semantic predictability-based reduction, clear speech, and theory of mind.

2.3 Experiment 1b: Semantic predictability-based reduction and clear speech

2.3.1 Introduction

The goals of this experiment were directly motivated by the preceding discussion. It sought to: replicate the unexpected interaction between AQ and word predictability; control for possible variance in IQ, and relate intelligence to variance in the AQ and RMITE scores; and to investigate interactions between semantic predictability-based reduction and clear speech, and ToM variation and clear speech. To this end, the experiment involved an extended version of the sentence reading task from experiment 1a, the AQ and RMITE tasks, and an IQ estimation task.

2.3.2 Methods

2.3.2.1 Participants

Twenty-five undergraduates participated in the experiment for partial course credit. All participants were monolingual American English speakers with no reported history of language, speech, or autism spectrum disorders. None of the participants had previously participated in experiment 1a.
2.3.2.2 Procedure

The same task and materials as in experiment 1a were used to elicit semantic predictability based reduction. Participants were instructed to read the sentences “as if talking to a friend”; sentences were presented on screen for 3.5 seconds at a time. The sentences were blocked by predictability condition, and randomized within each block. The order of blocks was counterbalanced between participants. After the first reading, participants were told that they would read the sentences again, but this time “as if talking to someone with a hearing impairment or who is a non-native speaker of English.” The sentences were presented for 4 seconds each in this condition,\(^8\) other aspects remained the same, although the exact order of the sentences was re-randomized. Following the reading task, participants completed the AQ questionnaire, the RMITE test, and an approximate IQ assessment (KBIT-2 Kaufman and Kaufman, 2004).

2.3.2.3 Measurements

As in experiment 1a, the word duration, vowel duration, and midpoint F1 and F2 of the target words were measured.

2.3.2.4 Analysis

A linear mixed effects regression model was constructed to model each of the acoustic variables—word duration, vowel duration, and vowel dispersion. Fixed effects were predictability condition (high versus low, reference: high), speech style (plain versus clear, reference: plain), AQ score, RMITE score, and estimated IQ score. Two-way interactions between predictability condition and speech style, predictability and each

\(^8\)Pilot testing revealed that participants found presentations of 3.5 seconds in the clear speech condition to be “too fast”. This lengthened time also implicitly encouraged the participants to speak clearly.
of the individual difference scores, and speech style and each of the individual difference scores were also entered. Finally, three-way interactions between predictability condition, speech style, and each of the individual difference variables were included. Therefore, the fixed effect structure was as follows:

\[
(\text{predictability} \times \text{style}) \times (\text{AQ} + \text{RMITE} + \text{IQ})
\]

All continuous variables were centered around the mean before being entered into the model. Random intercepts for talker and word identity were included, with random slopes for predictability condition and style for both talker and word. The calculation of \(p\)-values was the same as experiment 1a.

From a total of 4,100 tokens (41 words \(\times\) 2 predictability conditions \(\times\) 2 styles \(\times\) 25 talkers), 10 disfluent utterances and misreadings were excluded. Additionally, any measurements which were more than three standard deviations away from a subject’s mean were also removed, resulting in a total of 3,988 word duration measures, 3,978 vowel duration measures, and 3,971 vowel dispersion measures.

### 2.3.3 Results

#### 2.3.3.1 Individual differences scores

In this sample, AQ scores ranged from 82 to 127, with a mean of 107.80 (\(SD = 12.60\)); RMITE scores ranged from 19 to 36, with a mean of 27.84 (\(SD = 4.31\)); and estimated IQ scores ranged from 81 to 133, with a mean of 108.20 (\(SD = 13.97\)). These AQ and RMITE scores are comparable to those of experiment 1a in terms of range, central tendency, and dispersion. Table 2.10 shows a correlation matrix, with \(p\)-values, between these three variables. Although the correlation between IQ and RMITE scores appears
<table>
<thead>
<tr>
<th></th>
<th>AQ score</th>
<th>RMITE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMITE score</td>
<td>-.206 (.324)</td>
<td></td>
</tr>
<tr>
<td>IQ score</td>
<td>-.022 (.917)</td>
<td>.431 (.031)</td>
</tr>
</tbody>
</table>

Table 2.10: Correlation matrix of the individual difference scores in experiment 1b, also including p-values for the significance of each pairwise comparison.

significant at $\alpha = .05$, it was not significant after applying a Bonferroni correction for multiple comparisons ($\alpha = .05/3 \approx .017 < .031$).

Nevertheless, the finding of a significant correlation would not be surprising. A meta-analysis by Baker et al. (2014) of reports of RMITE and IQ scores revealed that general intelligence correlates positively with RMITE score ($r = .24 \pm .06$). This relationship may be due to the reliance on verbal labels in the task. Peterson and Miller (2012) found that individual differences in facial recognition ability could not predict RMITE scores, which underscores the role of language in completing the task (see also Johnston et al., 2008, for a critical review of the linking hypotheses underlying the RMITE task).

2.3.3.2 Word duration

Table 2.11 shows the model output for word duration in experiment 1b. Two significant simple effects were observed, those of word predictability and speech style. As expected, words in the LP condition were longer ($M = 443\text{ms}$) than those in the HP condition ($M = 433\text{ms}$), confirming that predictability-based reduction occurred. Additionally, words in the clear speech condition were significantly longer ($M = 483\text{ms}$).
Table 2.11: Output for the word duration model for experiment 1b.

than those in the plain speech condition \(M = 393\text{ms}\), consistent with decades of previous research on clear speech effects (Smiljanić and Bradlow, 2009).

Additionally, talker AQ score was observed to significantly interact with style, such that talkers with a higher AQ score produced shorter word durations in the clear speech condition relative to talkers with a lower AQ score. In other words, the higher AQ talkers had a smaller clear speech effect—their words were less enhanced, relative to the plain condition. AQ score additionally interacted with both style and word predictability in a significant three-way interaction. The effect of this interaction was that high AQ listeners significantly enhanced their distinction between LP and HP words—as in Experiment 1a—but only in the clear speech condition, and not in the plain speech.
condition. This effect is visualized in Figure 2.8; the interactive effect of AQ and style can also be observed.

![Word predictability](image)

Figure 2.8: Word duration as a function of talker AQ score, split by predictability condition. Left panel shows plain speech style, right panel clear speech.

### 2.3.3.3 Vowel duration

Table 2.12 shows the output for the vowel duration model in experiment 1b. As in the word duration model, word predictability was significant, such that words in the LP condition had longer vowels ($M = 220\text{ms}$) than words in the HP condition ($M = 211\text{ms}$). Likewise, speech style was significant, such that words in clear speech had longer vowels ($M = 238\text{ms}$) than words in plain speech ($M = 192\text{ms}$). No other significant effects were observed.
### Table 2.12: Output for the vowel duration model for experiment 1b.

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>$-27.201$</td>
<td>$-3.303$</td>
<td>$.001$ ***</td>
</tr>
<tr>
<td>Pred.: LP</td>
<td>$8.589$</td>
<td>$4.432$</td>
<td>$&lt;.001$ ***</td>
</tr>
<tr>
<td>Style: Clear</td>
<td>$44.838$</td>
<td>$8.122$</td>
<td>$&lt;.001$ ***</td>
</tr>
<tr>
<td>AQ</td>
<td>$-0.243$</td>
<td>$-0.499$</td>
<td>$.618$</td>
</tr>
<tr>
<td>RMITE</td>
<td>$-0.475$</td>
<td>$-0.300$</td>
<td>$.764$</td>
</tr>
<tr>
<td>IQ</td>
<td>$-0.154$</td>
<td>$-0.322$</td>
<td>$.748$</td>
</tr>
<tr>
<td>Pred. $\times$ Style</td>
<td>$0.754$</td>
<td>$0.323$</td>
<td>$.746$</td>
</tr>
<tr>
<td>Pred. $\times$ AQ</td>
<td>$0.205$</td>
<td>$1.316$</td>
<td>$.188$</td>
</tr>
<tr>
<td>Pred. $\times$ RMITE</td>
<td>$0.891$</td>
<td>$1.752$</td>
<td>$.080$</td>
</tr>
<tr>
<td>Pred. $\times$ IQ</td>
<td>$0.130$</td>
<td>$0.855$</td>
<td>$.393$</td>
</tr>
<tr>
<td>Style $\times$ AQ</td>
<td>$-0.618$</td>
<td>$-1.570$</td>
<td>$.116$</td>
</tr>
<tr>
<td>Style $\times$ RMITE</td>
<td>$0.447$</td>
<td>$0.350$</td>
<td>$.726$</td>
</tr>
<tr>
<td>Style $\times$ IQ</td>
<td>$0.331$</td>
<td>$0.860$</td>
<td>$.390$</td>
</tr>
<tr>
<td>Pred. $\times$ Style $\times$ AQ</td>
<td>$-0.111$</td>
<td>$-0.575$</td>
<td>$.565$</td>
</tr>
<tr>
<td>Pred. $\times$ Style $\times$ RMITE</td>
<td>$-0.259$</td>
<td>$-0.412$</td>
<td>$.681$</td>
</tr>
<tr>
<td>Pred. $\times$ Style $\times$ IQ</td>
<td>$-0.137$</td>
<td>$-0.725$</td>
<td>$.468$</td>
</tr>
</tbody>
</table>

2.3.3.4 **Vowel dispersion**

Table 2.13 shows the output for the vowel dispersion model in experiment 1b. Consistent with prior research on clear speech (e.g. Krause and Braida, 2004; Picheny et al., 1986), words in plain speech had less disperse vowels ($M = 2.735$ERB) than words in clear speech ($M = 3.018$ERB). A small but significant simple effect of word predictability also was observed: words in the HP condition had slightly less disperse vowels ($M = 2.868$ERB) than words in the LP condition ($M = 2.888$ERB). These two effects also interacted significantly such that the predictability effect was larger in the plain speech style than in the clear speech style ($M^\text{HP}_{\text{plain}} = 2.712$, $M^\text{LP}_{\text{plain}} = 2.758$, $M^\text{HP}_{\text{clear}} = 3.021$, $M^\text{LP}_{\text{clear}} = 3.016$), conceptually concordant with Baker and Bradlow’s
<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.178</td>
<td>-0.925</td>
<td>.355</td>
</tr>
<tr>
<td>Pred.: LP</td>
<td>0.059</td>
<td>2.166</td>
<td>.030 *</td>
</tr>
<tr>
<td>Style: Clear</td>
<td>0.327</td>
<td>7.411</td>
<td>&lt;.001 ***</td>
</tr>
<tr>
<td>AQ</td>
<td>-0.005</td>
<td>-0.962</td>
<td>.336</td>
</tr>
<tr>
<td>RMITE</td>
<td>0.014</td>
<td>0.899</td>
<td>.369</td>
</tr>
<tr>
<td>IQ</td>
<td>-0.001</td>
<td>-0.286</td>
<td>.775</td>
</tr>
<tr>
<td>Pred. × Style</td>
<td>-0.072</td>
<td>-2.354</td>
<td>.019 *</td>
</tr>
<tr>
<td>Pred. × AQ</td>
<td>0.002</td>
<td>1.013</td>
<td>.311</td>
</tr>
<tr>
<td>Pred. × RMITE</td>
<td>-0.009</td>
<td>-1.581</td>
<td>.114</td>
</tr>
<tr>
<td>Pred. × IQ</td>
<td>0.003</td>
<td>1.547</td>
<td>.122</td>
</tr>
<tr>
<td>Style × AQ</td>
<td>-0.003</td>
<td>-0.847</td>
<td>.397</td>
</tr>
<tr>
<td>Style × RMITE</td>
<td>-0.005</td>
<td>-0.481</td>
<td>.630</td>
</tr>
<tr>
<td>Style × IQ</td>
<td>0.000</td>
<td>0.115</td>
<td>.909</td>
</tr>
<tr>
<td>Pred. × Style × AQ</td>
<td>0.003</td>
<td>1.132</td>
<td>.258</td>
</tr>
<tr>
<td>Pred. × Style × RMITE</td>
<td>0.011</td>
<td>1.278</td>
<td>.201</td>
</tr>
<tr>
<td>Pred. × Style × IQ</td>
<td>-0.004</td>
<td>-1.787</td>
<td>.074</td>
</tr>
</tbody>
</table>

Table 2.13: Output for the vowel dispersion model for experiment 1b.

(2009) finding of larger mention effects on high-frequency words in plain speech relative to clear speech. No other effects or interactions were observed.

### 2.3.4 Discussion

The results of this experiment have, in general terms, replicated the effects observed in experiment 1a. Additionally, this experiment provided evidence that general intelligence could not account for the results: the IQ estimate failed to predict any acoustic variable or correlate with any ToM measure, supporting the conclusion from experiment 1a that variation in theory of mind is the relevant factor. Some inconsistencies between the two experiments, however, were observed. First, this experiment recorded
a small vowel dispersion effect due to predictability which was not observed in experiment 1a. The coefficient for this effect was small—0.059ERB—but significant. The reason for this difference is not known, although experiment 1b’s 25 participants against experiment 1a’s 21 represents a modest increase in power. In any case, a dispersion effect is consistent with previous literature on semantic predictability-based reduction (Scarborough, 2010).

The second inconsistency is that in this experiment, the interaction between predictability and ToM ability was only significant in the clear speech condition. Experiment 1a did not have a clear speech condition, and thus the speech can be regarded as plain. Indeed, the instructions for 1a were the same as those in the plain condition in 1b. A potential explanation for this curious effect is that the presence of the clear speech condition in 1b led the talkers to generally hypoarticulate their plain speech, which could wash away any effects of predictability. The instructions for 1b were very clear that talkers would have two blocks of speaking, the first “as if talking to a friend” and the second “as if talking to someone with a hearing impairment or who is a non-native speaker of English.” It is possible that hypoarticulation increased in the plain block of 1b in anticipation of the hyperspeech required for the clear block. In experiment 1a, on the other hand, there was no clear speech block and thus no anticipation. This tentative explanation finds some support in comparing the acoustic data of experiments 1a with 1b: a t-test comparing subject means suggests that the mean word duration ($M = 393\text{ms}$) from 1b’s plain speech block was significantly hypoarticulated relative to the word duration ($M = 420\text{ms}$, $t(35.562) = 2.179, p = .036$) from experiment 1a. Vowel dispersion, however, did not differ between 1b’s plain speech ($M = 2.732\text{ERB}$) and 1a ($M = 2.844\text{ERB}$, $t(36.843) = 1.233, p = .225$).

The clear speech manipulation itself also yielded interesting results. As expected, words in clear speech were longer and had longer and more disperse vowels
than words in plain speech. Additionally, talkers with higher AQ scores had smaller clear speech effects—these talkers did not enhance their speech as much as the talkers with lower AQ scores. This finding is a clear vindication of the role of theory of mind in listener-oriented speech styles: the talkers with poor ToM (high AQ scores) were less skilled at adapting their speech for their (imagined) interlocutor than the talkers with good ToM. Finally, the fact that predictability effects were observed in the clear speech in addition to the plain speech suggests that these effects are well outside of conscious control. If the explicit control of clear speech were able to ‘override’ the influence of predictability on production—that is, if talkers could engage and disengage the predictability effects at will—then we would expect to observe great variability in clear speech, including the possibility of the absence of predictability effects. The data do not match this prediction, and therefore, there is no evidence that these predictability effects are subject to conscious control.

2.4 Experiment 2: Word-specific VOT effects

2.4.1 Introduction

This experiment examined relationships among word-specific VOT effects and variability in theory of mind.

2.4.2 Methods

2.4.2.1 Participants

Nineteen undergraduates participated in the experiment for partial course credit. All participants were monolingual American English speakers with no reported history of
language, speech, or autism spectrum disorders. None of the participants had previously participated in experiments 1a or 1b.

2.4.2.2 Materials

Stimulus words for this experiment were taken from both Baese-Berk and Goldrick's (2009) and Fox et al.'s (2015) studies. Lexical frequency and neighborhood density values for the stimuli were taken from the Hoosier Mental Lexicon (Nusbaum et al., 1984). Words which were not present in the Hoosier Mental Lexicon were not included in the experiment. Stimuli varied in whether they had an initial voicing minimal pair neighbor or not. For example, cod has a minimal pair neighbor in god, but cop does not have such a neighbor (*gop). A total of 42 stimuli items had minimal pair neighbors, while 50 did not. A full list of the stimuli, including fillers and presentation order, is presented in Appendix D.

2.4.2.3 Procedure

The participant sat across a table from a confederate in a double-walled sound attenuated booth. Both were seated in front of a computer screen, and they could not see each other’s screens. Both wore head-mounted microphones. On the participant’s screen, three words appeared. After 1000ms, the target word turned red, thereby becoming highlighted. The participant then instructed the confederate to click on the highlighted word. The participant was given no direct orders in how to phrase their instruction. The frame of “click on the shoe” was suggested, but it was mentioned that the participant was free to speak in any way they pleased. The confederate moved their mouse and clicked, and after a random interval of 500 to 2000ms, the participant’s screen advanced to the next trial. After the interactive task, the confederate revealed to the
participant that she was an experimenter and not a fellow participant and then left the booth after some debriefing questions. Next, the participant completed the AQ questionnaire, the RMITE test, and the IQ estimation.

The confederate was an undergraduate research assistant. Various efforts were made to maintain the illusion that the confederate was in fact a fellow participant and not an experimenter. For example, the confederate went through the same consent procedure as the participant; similarly, the confederate, like the participant, wore a head-mounted mic and their speech was ostensibly recorded. Due to a scheduling conflict, one participant was run with the experimenter (the author) acting as a confederate rather than the research assistant; these data are included in the present analysis as they did not appear to be different in any way from the rest of the data.

Each trial was in one of three conditions: a “no minimal pair” condition, where the target word was like *cop* and doesn’t have a minimal pair competitor; a “minimal pair without competitor” condition, where the target word was like *cod* and was presented with two other filler words; and a “minimal pair with competitor” condition, where the target word was like *cod* and was presented alongside its competitor *god* and one filler. Examples are provided in Table 2.14. Each participant saw each target word only once; therefore, the manipulation of what context the minimal pair words were presented in was between subjects, while the manipulation of whether or not the

<table>
<thead>
<tr>
<th>Condition</th>
<th>Target</th>
<th>Competitor</th>
<th>Distractor(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No minimal pair</td>
<td>pork</td>
<td>soil, brick</td>
<td></td>
</tr>
<tr>
<td>Minimal pair without competitor</td>
<td>punk</td>
<td>date, league</td>
<td></td>
</tr>
<tr>
<td>Minimal pair with competitor</td>
<td>punk</td>
<td>bunk</td>
<td>league</td>
</tr>
</tbody>
</table>

Table 2.14: Example stimulus words presented to participants in each of the three conditions of experiment 2.
target had a minimal pair was within subjects. Of the 92 total trials, 50 were in the “no minimal pair” condition, since those stimuli did not have a minimal pair. The remaining 42 stimuli were balanced into two lists, where each stimulus item appeared in the “minimal pair without competitor” in one list and in the “minimal pair with competitor” condition in the other list. Therefore, each participant saw all 92 stimuli exactly once in a combination of the three conditions. See Appendix D for a list of stimuli, distractors, and fillers, and the makeup of the lists.

2.4.2.4 Measurements

Voice onset time of the word-initial voiceless stops in target words was automatically measured using Sonderegger and Keshet’s (2012) AutoVOT software. These measurements were then hand-corrected.

2.4.2.5 Analysis

From a total of 1,748 target tokens (19 participants × 92 targets), tokens with disfluent productions were removed. VOT values more than three standard deviations away from a subject’s mean were also removed. Thus, a total of 1,735 data points were available for analysis.

Since there were so many variables of interest, and many potentially interesting interactions, the construction of one grand model with many effects and interactions would likely overfit the data. Therefore, the modeling procedure for this experiment was different from that used in experiments 1a and 1b. Instead, a forward stepwise procedure was followed, beginning with a model with only an intercept as a fixed effect, and random intercepts of talker and word. Effects of stop place (labial, alveolar, velar, baseline: alveolar), context (no minimal pair, minimal pair with no competitor,
minimal pair with competitor; coded with Helmert contrasts), neighborhood density, log frequency, talker AQ score, talker RMITE score, and talker IQ estimate were added one at a time and compared via likelihood ratio testing to the baseline model. The fixed effect that contributed the most to the data likelihood was selected and added as a fixed effect and, when appropriate, as a random slope. The procedure then continued to test and add fixed effects. Once more than one fixed effect was established in the model, the interactions between that effect and all other effects previous in the model were similarly evaluated and retained if significant. Interactions were never added to random slopes due to the danger of overfitting. All continuous variables were centered prior to being entered into the model.

2.4.3 Results

In this sample, AQ scores ranged from 90 to 126, with a mean of 112.47 (SD = 10.80); RMITE scores ranged from 24 to 36, with a mean of 29.32 (SD = 2.98); and IQ estimates ranged from 93 to 140, with a mean of 113.84 (SD = 11.65). These scores are similar to those of experiments 1a and 1b in terms of range, central tendency, and dispersion. Table 2.15 shows a correlation matrix, with p-values, between these three variables. As can be seen, none of the correlations were significant.

The first iteration of the model selection procedure added the effect of stop place ($\chi^2(2) = 60.024$, $p < .001$). The second iteration failed to find any significant data likelihood improvement by adding any factors, and thus the final model simply involves the fixed effect of consonant place. Table 2.16 shows the model summary. As expected, velar stops had significantly longer VOT values ($M = 88.3$ms) than alveolar stops.

---

9An example of an inappropriate random slope would be a slope of frequency as a word-level random effect, or a slope of AQ as a talker-level random effect.
Table 2.15: Correlation matrix of the individual difference scores in experiment 2, also including p-values for the significance of each pairwise comparison.

<table>
<thead>
<tr>
<th></th>
<th>AQ score</th>
<th>RMITE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$(p)$</td>
<td>$r$</td>
<td>$(p)$</td>
</tr>
<tr>
<td>RMITE score</td>
<td>.095</td>
<td>(.699)</td>
<td>.233</td>
<td>(.336)</td>
</tr>
<tr>
<td>IQ score</td>
<td>−.001</td>
<td>(.998)</td>
<td>.233</td>
<td>(.336)</td>
</tr>
</tbody>
</table>

Table 2.16: Output for the optimal model predicting VOT from the experiment 2 results.

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.926</td>
<td>0.572</td>
<td>.567</td>
</tr>
<tr>
<td>Place: Labial</td>
<td>−9.012</td>
<td>−4.841</td>
<td>&lt;.001  ***</td>
</tr>
<tr>
<td>Place: Velar</td>
<td>3.808</td>
<td>2.237</td>
<td>.025   *</td>
</tr>
</tbody>
</table>

stops ($M = 84.4$ms), which were in turn significantly longer than labial stops ($M = 75.3$ms).

2.4.4 Discussion

The lack of any effects beyond that of stop place is somewhat surprising. Baese-Berk and Goldrick’s (2009) finding has been replicated on several occasions (Bullock-Rest et al., 2013; Fox et al., 2015; Goldrick et al., 2013; Kirov and Wilson, 2012; Peramunage et al., 2011). Figure 2.9 shows the (lack of) effect of the experimental manipulations in the current study, with the reported mean VOTs from Baese-Berk and Goldrick’s (2009) study indicated by open circles. While Baese-Berk and Goldrick’s (2009) data show a clear trend, there is no difference between the conditions for the present data.
One potential source of this difference is the materials used. Fox et al. (2015) used stimuli carefully balanced for neighborhood density, and similarly did not find a minimal pair effect. Instead, however, they found a neighborhood density effect, whereby words with more neighbors had longer VOT values. This effect is also absent from the present data, although neighborhood density provided a marginally significant improvement to data likelihood ($\chi^2(1) = 2.920, p = .088$). This trend is visualized in Figure 2.10. However, this trend may be an epiphenomenon of the stimuli used—words with velar stops had a higher mean neighborhood density (22.79) than alveolar (17.71) or labial (17.31) stops, and velar stops, as mentioned, tend to have longer VOT values than other stop places.

Although drawing conclusions from null results can be problematic, the lack of any effects or interactions with the individual difference scores of AQ, RMITE, or IQ suggest that phonetic variation due to density is unrelated to these factors. This
conclusion will not be surprising to some: of all of the parameters of phonetic reduction explained through listener-oriented accounts, neighborhood density is perhaps the most problematic. Although the relationship between semantic predictability and listener expectations is relatively clear, the connection between neighborhood density and listener expectation is somewhat more tenuous. It is not clear if a word’s being in a dense neighborhood means it is difficult for a listener to access, due to competition from neighbors, or if it is easier to access, due to the relatively high phonotactic probability of the component phoneme sequences. The next experiment studies more broadly the influences of lexical neighborhood and frequency on word production, and its interactions with individual differences in theory of mind.

Figure 2.10: Voice onset time as a function of neighborhood density.
2.5 Experiment 3: Lexical neighborhood density and frequency

2.5.1 Introduction

This experiment examined the effects of neighborhood density and frequency on reduction, and the interactions with individual variation in theory of mind.

2.5.2 Methods

2.5.2.1 Materials

Target words consisted of 236 English words drawn from a set of stimulus paragraphs developed by Clopper and colleagues (Burdin and Clopper, 2015; Burdin et al., 2014a,b). Words were selected to balance both frequency and neighborhood density as recorded in the Hoosier Mental Lexicon (Nusbaum et al., 1984). All words but two were monosyllabic (the two exceptions were disyllabic). The stressed vowels in each stimulus were /æ/ (42 words), /ɛ/ (39 words), /i/ (45 words), /ɑ/ (33 words), /ɔ/ (39 words), and /u/ (38 words). The stimuli are listed in full in Appendix E.

2.5.2.2 Procedure

The experiment took place in a double-walled sound attenuated booth. Words were presented individually on a computer screen and participants were instructed to read each word aloud. Their speech was recorded via a head-mounted microphone. After reading a word, participants pressed any key on the computer to advance to the next trial. After the word-reading, participants completed the AQ, RMITE, and IQ tasks.
2.5.2.3 Measurements

A first-pass alignment of the words was conducted using the Penn Forced Aligner (Yuan and Liberman, 2008). These alignments were subsequently hand-checked and corrected. Measurements of word duration, vowel duration, and vowel dispersion were calculated as in experiment 1a.

2.5.2.4 Participants

Eighteen participants completed the task for partial course credit. Eligibility was restricted to native monolingual speakers of American English with no history of speech, language, or hearing disorders. None of the participants reported any history of autism-spectrum conditions.

2.5.2.5 Analysis

Mispronunciations, disfluencies, and restarted utterances were excluded from the analysis, resulting in the removal of 31 tokens for a total of 4,217 analyzable tokens. Twenty-two tokens had incorrect formant measures and were excluded prior to the calculation of the dispersion metric. Additionally, all formant values more than three standard deviations away from their talker by-vowel means were excluded, as were all duration values more than three standard deviations away from their talker means. This process resulted in a total of 4,195 dispersion measures, 4,202 word duration measures, and 4,193 vowel duration measures being entered into the analysis.

Linear mixed effects models were constructed for word duration, vowel duration, and vowel dispersion. Word log frequency, neighborhood density, talker AQ, and talker RMITE score were added as fixed effects. Two-way interactions between frequency and density, frequency and AQ, frequency and RMITE, density and AQ, and
density and RMITE were also included. Finally, talker IQ and the number of phonemes in the target word were included as covariates. A random intercept for talker identity was added with random slopes of frequency, density, and number of phonemes, as well as a random intercept of word with random slopes of AQ, RMITE, and IQ score. As in the previous experiments, $p$-values were calculated by treating the absolute $t$-statistic as if it came from a $t$-distribution with degrees of freedom equal to the number of data points minus the number of model parameters.

2.5.3 Results

2.5.3.1 Individual difference scores

Due to equipment failure one participant did not fully complete the IQ estimation task, and therefore does not have an IQ estimate. In this sample, AQ scores ranged from 86 to 126, with a mean of 110.22 ($SD = 11.23$); RMITE scores ranged from 26 to 35, with a mean of 30.72 ($SD = 2.70$); and estimated IQ scores ranged from 97 to 124, with a mean of 111.18 ($SD = 7.37$). These scores are similar to those of experiments 1a, 1b, and 2 in terms of range, central tendency, and dispersion. Table 2.17 shows a correlation matrix, with $p$-values, between these three variables. As can be seen, none of these variables were significantly correlated with each other.

2.5.3.2 Word duration

Table 2.18 shows the model output for the word duration model in experiment 3. A significant effect of IQ was observed, such that talkers with a higher estimated IQ tended to have longer words—they spoke more slowly. An effect of neighborhood density was also observed, such that words with more neighbors tended to have shorter duration, even after controlling for the number of phonemes in the word. A visualization of this
<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.960</td>
<td>0.079</td>
<td>.937</td>
</tr>
<tr>
<td>IQ</td>
<td>5.098</td>
<td>3.752</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Number of phonemes</td>
<td>7.717</td>
<td>0.871</td>
<td>.384</td>
</tr>
<tr>
<td>Log frequency</td>
<td>-8.903</td>
<td>-1.533</td>
<td>.125</td>
</tr>
<tr>
<td>Density</td>
<td>-5.694</td>
<td>-7.574</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>AQ</td>
<td>0.580</td>
<td>0.604</td>
<td>.546</td>
</tr>
<tr>
<td>RMITE</td>
<td>-0.540</td>
<td>-0.124</td>
<td>.901</td>
</tr>
<tr>
<td>Log frequency $\times$ Density</td>
<td>0.589</td>
<td>0.886</td>
<td>.376</td>
</tr>
<tr>
<td>Log frequency $\times$ AQ</td>
<td>-0.034</td>
<td>-0.240</td>
<td>.811</td>
</tr>
<tr>
<td>Log frequency $\times$ RMITE</td>
<td>-0.769</td>
<td>-1.149</td>
<td>.251</td>
</tr>
<tr>
<td>Density $\times$ AQ</td>
<td>-0.026</td>
<td>-0.864</td>
<td>.388</td>
</tr>
<tr>
<td>Density $\times$ RMITE</td>
<td>-0.031</td>
<td>-0.228</td>
<td>.820</td>
</tr>
<tr>
<td>Log frequency $\times$ Density $\times$ AQ</td>
<td>-0.010</td>
<td>-0.671</td>
<td>.502</td>
</tr>
<tr>
<td>Log frequency $\times$ Density $\times$ RMITE</td>
<td>-0.137</td>
<td>-2.095</td>
<td>.036 *</td>
</tr>
</tbody>
</table>

Table 2.18: Output for word duration model, experiment 3.

effect is shown in Figure 2.11. Additionally, a three-way interaction between talker RMITE, word frequency, and number of neighbors was observed. Since this effect involves the interaction of three continuous variables on the dependent variable, which is also continuous, graphing this interaction is difficult. An attempt has been made in Figure 2.12. To make this figure, frequency was split into two categories, high (log frequency > 2.5) and low; density was split into dense neighborhoods (> 11 neighbors)
and sparse neighborhoods. The left panel shows the pattern of results in words in dense neighborhoods; the right panel shows the pattern for the words in sparse neighborhoods. As can be seen in the right panel, the separation between the two trend lines indicates a robust frequency effect: high-frequency words (in sparse neighborhoods) are shorter in duration ($M = 543\text{ms}$) than low-frequency words ($M = 563\text{ms}$).

In the left panel, we observe the interaction with RMITE score. In words with dense neighborhoods, for talkers with high RMITE scores, the same frequency effect can be observed: high-frequency words are shorter than low-frequency words. However, as RMITE score decreases, this effect diminishes, until, for the participants with the lowest RMITE scores, there is no difference in duration between the high- and low-frequency words with dense neighborhoods.

![Figure 2.11: Word duration as a function of number of lexical neighbors. Points indicate means of individual word types, and bars the standard error.](image)

Figure 2.11: Word duration as a function of number of lexical neighbors. Points indicate means of individual word types, and bars the standard error.
Figure 2.12: Word duration as a function of talker RMITE score, split by word frequency and neighborhood density. See text for details on binning.

### 2.5.3.3 Vowel duration

Table 2.19 shows the model output for the vowel duration model in experiment 3. A simple effect of number of phonemes was observed: words with more phonemes tended to have shorter vowels than words with fewer phonemes ($M_2 = 306\text{ms}$, $M_3 = 229\text{ms}$, $M_4 = 218\text{ms}$, $M_5 = 200\text{ms}$). This effect is consistent with previous research which suggests that syllables with more consonants tend to have shorter vowels than syllables with fewer consonants (Schwarzlose and Bradlow, 2001). This equivalence between number of phonemes and number of consonants per syllable can be drawn because the stimuli involved in this experiment monosyllables, with the exception of *tunic* and *water*. Additionally, a simple effect of neighborhood density was observed,
Table 2.19: Output for vowel duration model, experiment 3.

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.167</td>
<td>0.142</td>
<td>.887</td>
</tr>
<tr>
<td>IQ</td>
<td>0.948</td>
<td>1.266</td>
<td>.206</td>
</tr>
<tr>
<td>Number of phonemes</td>
<td>-34.350</td>
<td>-5.728</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Log frequency</td>
<td>-4.772</td>
<td>-1.279</td>
<td>.201</td>
</tr>
<tr>
<td>Density</td>
<td>-1.960</td>
<td>-4.011</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>AQ</td>
<td>0.351</td>
<td>0.496</td>
<td>.620</td>
</tr>
<tr>
<td>RMITE</td>
<td>-0.906</td>
<td>-0.280</td>
<td>.779</td>
</tr>
<tr>
<td>Log frequency × Density</td>
<td>0.345</td>
<td>0.791</td>
<td>.429</td>
</tr>
<tr>
<td>Log frequency × AQ</td>
<td>-0.084</td>
<td>-1.229</td>
<td>.219</td>
</tr>
<tr>
<td>Log frequency × RMITE</td>
<td>-0.474</td>
<td>-1.496</td>
<td>.135</td>
</tr>
<tr>
<td>Density × AQ</td>
<td>0.012</td>
<td>1.064</td>
<td>.288</td>
</tr>
<tr>
<td>Density × RMITE</td>
<td>0.010</td>
<td>0.193</td>
<td>.847</td>
</tr>
<tr>
<td>Log frequency × Density × AQ</td>
<td>0.002</td>
<td>0.193</td>
<td>.847</td>
</tr>
<tr>
<td>Log frequency × Density × RMITE</td>
<td>0.011</td>
<td>0.297</td>
<td>.767</td>
</tr>
</tbody>
</table>

whereby words with more neighbors tended to have shorter vowels than words with fewer neighbors. No other significant effects or interactions were observed.

2.5.3.4 Vowel dispersion

Table 2.20 shows the model output for the vowel dispersion model. Only one significant effect was observed, which was of number of phonemes. Words with more phonemes tended to have less disperse vowels than words with fewer phonemes ($M_2 = 3.777ERB, M_3 = 3.257ERB, M_4 = 3.063ERB, M_5 = 2.884ERB$). This effect is consistent with the vowel duration shortening for words with more consonants, and is expected given the relationship between vowel duration and vowel dispersion (Moon and Lindblom, 1994).
<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.011</td>
<td>-0.092</td>
<td>.927</td>
</tr>
<tr>
<td>IQ</td>
<td>0.017</td>
<td>1.363</td>
<td>.173</td>
</tr>
<tr>
<td>Number of phonemes</td>
<td>-0.346</td>
<td>-2.287</td>
<td>.022 *</td>
</tr>
<tr>
<td>Log frequency</td>
<td>-0.142</td>
<td>-1.406</td>
<td>.160</td>
</tr>
<tr>
<td>Density</td>
<td>-0.011</td>
<td>-0.957</td>
<td>.338</td>
</tr>
<tr>
<td>AQ</td>
<td>0.001</td>
<td>0.093</td>
<td>.926</td>
</tr>
<tr>
<td>RMITE</td>
<td>0.017</td>
<td>0.442</td>
<td>.658</td>
</tr>
<tr>
<td>Log frequency × Density</td>
<td>-0.001</td>
<td>-0.101</td>
<td>.920</td>
</tr>
<tr>
<td>Log frequency × AQ</td>
<td>0.000</td>
<td>0.167</td>
<td>.867</td>
</tr>
<tr>
<td>Log frequency × RMITE</td>
<td>-0.003</td>
<td>-0.510</td>
<td>.610</td>
</tr>
<tr>
<td>Density × AQ</td>
<td>0.000</td>
<td>1.184</td>
<td>.236</td>
</tr>
<tr>
<td>Density × RMITE</td>
<td>0.001</td>
<td>1.598</td>
<td>.110</td>
</tr>
<tr>
<td>Log frequency × Density × AQ</td>
<td>0.000</td>
<td>0.126</td>
<td>.899</td>
</tr>
<tr>
<td>Log frequency × Density × RMITE</td>
<td>0.000</td>
<td>-0.332</td>
<td>.740</td>
</tr>
</tbody>
</table>

Table 2.20: Output for vowel dispersion model, experiment 3.

### 2.5.4 Discussion

An effect of frequency on word duration was observed, with more frequent words being shorter in duration than less frequent words. However, this effect was significantly diminished for talkers with poor theory of mind for words in dense neighborhoods. This effect of theory of mind on reduction is in the opposite direction to that observed in experiments 1a and 1b. In those experiments, talkers with worse theory of mind had a larger extent of reduction between high- and low-predictability words than talkers with better theory of mind. Here, however, the generalization is the opposite: worse theory of mind leads to less reduction, as predicted by the listener-oriented account.

Neighborhood density was observed to be negatively correlated with word duration and vowel duration, such that words with more neighbors were in general shorter, and had shorter vowels, than words with fewer neighbors. No effect was observed
for vowel dispersion. This effect is unexpected for two principal reasons. First, most studies of neighborhood density have found effects in the spectral domain (i.e. vowel dispersion) but not the temporal domain. Although some studies have reported effects on duration (Burdin and Clopper, 2015; Kryuchkova and Tucker, 2012; Scarborough, 2010), others have failed to observe such effects (Munson and Solomon, 2004). Since more disperse vowels tend to require a longer time to produce than less disperse vowels (Lindblom, 1963), it has been posited that temporal and spectral effects go hand-in-hand (Flemming, 2010; Moon and Lindblom, 1994). Second, the majority of studies have found positive correlations between acoustic prominence and neighborhood density. That is, words with more neighbors tend to have longer and more disperse vowels (Burdin and Clopper, 2015; Wright, 2004) than words with fewer neighbors.

However, this unexpected result is concordant with those of Gahl et al. (2012), who found decreased vowel duration and dispersion in words with higher neighborhood density relative to words with lower neighborhood density, after controlling for a number of factors known to influence vowel duration and dispersion. In addition to the presence of a large number of control factors which many previous studies have neglected, Gahl et al. (2012) argued that the disparity between their results and the previous literature was due to their dataset being spontaneous speech, rather than the word list or read speech approach which typifies much of the literature. Subsequent work by Gahl (2015), however, demonstrated that a reanalysis of Wright's (2004) word list data that controls for consonantal content does not require neighborhood density to predict vowel dispersion. Taken together, the results obtained by Gahl et al. (2012) and Gahl (2015) suggest that great care should be exercised in the modeling of neighborhood density effects, and that once other factors are controlled for, a negative correlation between density and phonetic substance obtains.
The present data, then, provide an interesting case. These data show that neighborhood density negatively correlates with word duration, consistent with Gahl et al.’s (2012) results. The fact that consonantal context was not directly controlled for, and a negative result still obtained, could be interpreted as a serendipitous happenstance. Curiously, though, the word stimuli used were the same as those used by Burdin and Clopper (2015), and the result is opposite to theirs. Interpretation of Burdin and Clopper’s (2015) model coefficients suggest that for every extra lexical neighbor, words were on average 0.9ms longer. The word duration model in experiment 3 of the current study suggests that for every extra lexical neighbor, words were on average 5.7ms shorter. This discrepancy cannot be ascribed to stimulus differences since the stimuli were the same between both experiments. The main methodological difference between experiment 3 and Burdin and Clopper’s (2015) study was that experiment 3 involved participants reading words from a computer screen one at a time, whereas Burdin and Clopper’s (2015) experiment had participants read entire paragraphs. See also Chapter 4 for more detailed statistical comparison of Burdin and Clopper’s data (Burdin and Clopper, 2015; Burdin et al., 2014a,b) and the present data.

2.6 General Discussion

2.6.1 Summary of experiments

The experiments reported in this chapter investigated the relationship between phonetic reduction and individual variation in theory of mind. Experiment 1a established that the AQ and RMITE scores were useful measures of ToM for accounting for variance in perceptual accommodation to coarticulation (replicating Yu, 2010). Additionally, the experiment investigated second mention reduction and semantic predictability-based
reduction. Second mention reduction was observed in both the temporal and spectral domains: second mentions of words were shorter with less disperse vowels than first mentions. No interactions with ToM skill was observed. Semantic predictability-based reduction was also observed, but in the temporal domain only. An interaction with ToM skill was observed such that talkers with poorer ToM produced larger magnitudes of reduction—that is, their acoustic differences between predictable and unpredictable words was larger than the difference produced by talkers with better ToM.

Experiment 1b partially replicated this interaction, although statistical significance was reached only in the ‘clear speech’ condition, where participants were talking as if to someone with a hearing impairment or a non-native speaker of English. The data from experiment 1b also showed a small but significant effect of semantic predictability on vowel dispersion, such that vowels in predictable words were less disperse than vowels in unpredictable words, consistent with the temporal effects. Additionally, experiment 1b controlled for general intelligence, a possible confound on ToM skill.

Experiment 2 investigated word-specific VOT effects. Despite following Baese-Berk and Goldrick’s (2009) methods in many respects, this experiment failed to replicate either the minimal pair effects reported by Baese-Berk and Goldrick (2009) or the neighborhood density effects reported by Fox et al. (2015). The only significant determinant of VOT duration was place of articulation, an expected and unsurprising result (Lisker and Abramson, 1964).

Experiment 3 investigated effects of neighborhood density and frequency on phonetic reduction. Effects of neighborhood density were observed, such that words with more neighbors were shorter than words with fewer neighbors. Similarly, effects of frequency were observed, whereby more frequent words were shorter than less frequent words—but crucially, this effect was blocked for words in dense neighborhoods.
spoken by talkers with low RMITE scores. Those words were not affected by frequency, and the frequent and infrequent words were of similar durations.

2.6.2 Phonetic reduction

The majority of phonetic reduction processes observed here are relatively uncontroversial, and the results are expected given the literature. It is expected that highly frequent words are shorter than less frequent words (Gahl, 2008); that second mentions are reduced relative to first mentions (Baker and Bradlow, 2009); and that highly predictable words are shorter than unpredictable words (Scarborough, 2010; Turnbull and Clopper, 2013). However, the effects relating to neighborhood density are potentially more problematic, in that higher-density words exhibited more reduction than low-density words, in contrast to prior work (e.g. Wright, 2004).

From a developmental standpoint, the confluence of phonetic reduction on high-frequency and high-density words is not difficult to explain. As the learner (a child) develops her lexicon, a phonological system is constructed with Gestalt-like representations of productions, where whole syllables or sequences of phonemes are represented rather than individual phonemes (Ferguson, 1986; Ferguson and Farwell, 1975). Under the assumption that distinct motor scores are not multiply represented (Lindblom, 1992), it stands to reason that the learner acquires more experience and practice with the common production sequences. Production sequences can be more common by belonging to a frequent words, and by belonging to many words at the same time: in other words, the learner obtains more practice with sequences from frequent words than infrequent words, and from words in dense neighborhoods than words in sparse neighborhoods. Assuming that practice leads to fluency of production and more tight temporal control of articulators, and therefore decreases in segment duration, such a
framework provides a developmental mechanism for high-frequency and high-density words to exhibit temporal reduction relative to low-frequency and low-density words.

### 2.6.3 Individual differences

Experiments 1a, 1b, and 3 all revealed interactions between individual differences in theory of mind skill and phonetic reduction. Experiments 1a and 1b established that the extent of semantic predicability-based reduction varies as a function of ToM skill: talkers with poorer ToM had a larger durational difference between their productions of high- and low-predictability words than talkers with better ToM. Some of the participants with stronger theory of mind in fact did not produce any difference at all between the high- and low-predictability words. This result could be interpreted as a case of better pattern-recognition and systemizing skills among the lower-ToM participants, consistent with research on the broader autism phenotype (Baron-Cohen, 2008, 2009; Baron-Cohen et al., 2002). However, such a conclusion could be premature. Second mention reduction showed no relationship with the ToM measures; second mention reduction does not appear to vary as a function of theory of mind, suggesting that this kind of reduction arises from a process or mechanism that is cognitively distinct from that which leads to semantic predictability-based reduction. A pattern-recognition and systemizing account would predict that second mention reduction ought to follow the same patterns as semantic predictability-based reduction.

Experiment 3 established that lexical frequency effects interact with ToM skill. While talkers with good ToM skills consistently produced high frequency words with less phonetic substance than low frequency words, talkers with poor ToM skills did not produce this distinction consistently for high-density words. In other words, low density words exhibited frequency effects for all talkers; high-density words exhibited
frequency effects only for talkers with good ToM skills. Table 2.21 summarizes these effects for these four types of reduction examined.

<table>
<thead>
<tr>
<th>Reduction type</th>
<th>Interactions</th>
<th>More ToM leads to...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic predictability</td>
<td>✓</td>
<td>less reduction</td>
</tr>
<tr>
<td>Second mention</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Neighborhood density</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>✓</td>
<td>more reduction</td>
</tr>
</tbody>
</table>

Table 2.21: Summary of the four reduction types examined in experiments 1a, 1b, and 3, and their interactions with individual differences in theory of mind.

2.6.4 Listener-oriented theories

Turning now to the listener-oriented theories which form the theoretical basis for this study, it is possible to ask what implications the present results hold for these accounts of phonetic reduction. In the introduction, we outlined the major predictions of a strong interpretation of listener orientation: that ToM should negatively correlate with the extent and consistency of reduction. The data from the current experiments suggest that ToM negatively correlates with the extent of reduction due to frequency, for some words: talkers with poorer ToM had a smaller difference between high and low frequency words than talkers with better ToM. This result is consistent with a listener-oriented account: the talkers with poor ToM are less able to model their interlocutors and thus less adept at performing reduction at the appropriate times.

However, ToM failed to correlate with effects of neighborhood density and second mention reduction. These results are incompatible with the listener-oriented account, although a weaker version of the theory which allows for a ‘generic listener’ could accommodate these data (Turk, 2010). These results are also consistent with
talker-based or ‘egocentric’ accounts of speech production (Bard et al., 2000; Bard and Aylett, 2005; Keysar and Barr, 2005; Keysar et al., 2000), where phonetic reduction is a function of activation or accessibility within the speech production system. Words with many neighbors and words that are accessibly in the discourse context are highly activated, and therefore are produced in a reduced way, regardless of the listener’s needs.

Nevertheless, neither a weak listener-oriented theory nor an egocentric theory predict the ToM interactions observed for word frequency; additionally, neither kind of listener-oriented theory nor an egocentric theory are able to account for the effects of ToM on semantic predictability-based reduction observed in experiment 1a and 1b. Here, participants with poorer ToM exhibited a larger difference between their productions of high- and low-predictability words than did participants with stronger ToM. It is reasonable to assume that some of this difference is due to idiosyncratic individual variability; after all, it is well-established that some people have generally more intelligible and clearer speech than other people, regardless of context (Ferguson, 2004; Picheny et al., 1985). While some of this variance in speech production could be due to non-pathological physiological factors (e.g. mild dental malocclusions, Kummer, 2008), presumably a large aspect of this variance is due to cognitive and personality factors. There is currently no role for these factors in talker-oriented or listener-oriented theories.

## 2.7 Conclusion

This chapter described four experiments which examined relationships between phonetic reduction and individual variation in theory of mind. VOT was not observed to vary as a function of neighborhood density or target word minimal pair status. Second
mention reduction and neighborhood density effects were not found to vary with the ToM of the talker. Frequency effects on word duration were smaller for talkers with poorer ToM, consistent with the predictions of a listener-oriented theory. However, effects of semantic predictability were larger for talkers with poorer ToM, a result not predicted by either listener-oriented or talker-oriented theories. Taken together, these results suggest that the listener-oriented account of phonetic reduction enjoys only limited explanatory adequacy. Further, models of phonetic reduction may need to consider individual variation in cognition in order to adequately account for the observed effects.
Individual differences in the perception of phonetically reduced speech

3.1 Introduction

The 'listener-oriented' account of phonetic reduction (e.g. Aylett and Turk, 2004, 2006; Lindblom, 1990) posits that speech production is a balance between speaking clearly so that the interlocutor will understand, and reducing articulatory effort to conserve the talker's energy. This reasoning has been applied to explain the observed trends of contextually predictable words being phonetically reduced relative to contextually unpredictable words (e.g. Gahl and Garnsey, 2004; Lieberman, 1963; Raymond et al., 2006; Tily and Kuperman, 2012). If the word is predictable in context, the context provides information to the listener in addition to the perceived acoustic signal. Therefore, since the context boosts the overall likelihood of comprehension, the talker is free to reduce articulatory effort by producing the word with phonetic reduction (realized through durational shortening, segment deletion, vowel centralization, and a host of other acoustic parameters; see Clopper and Turnbull, 2015, for review). For words in unpredictable contexts, since the context provides no clues or very few clues for the listener, the talker must produce the word in an unreduced manner to ensure comprehension.
An essential assumption of this account is that speech which is not reduced is beneficial to the listener in some way, or that reduced speech is harder for the listener to process. If this assumption were not true, then there would be no imperative for the talker to speak in an unreduced manner. However, the exact formulation of what ‘beneficial’ means is not clearly articulated by these theories. It could refer to subjective judgements of speech as sounding more or less clear; speech intelligibility; speed and ease of lexical retrieval; or even speed and ease at a higher level of processing such as semantic judgements. This chapter tests this assumption and examines these various possible dimensions of ‘benefit’ through a series of speech perception experiments.

3.1.1 Effects of reduction in perception

It has been known for more than half a century that some kinds of reduced speech are less intelligible than unreduced speech. Lieberman (1963) found that words produced in a semantically predictable context—such as the last word in (1)—are less intelligible in isolation than words produced in a semantically less predictable context—such as the last word in (2). Similarly, words with reductions arising from repeated mention (‘second mention reduction’) have been shown to be less intelligible in isolation than first mentions (Bard and Anderson, 1994; Bard et al., 2000; Fowler and Housum, 1987). Indeed, some words can be so reduced as to be wholly unintelligible in isolation, requiring presentation in their original context in order for meaning to be recoverable (Ernestus et al., 2002).

(1) A stitch in time saves nine.

(2) The number that you will hear is nine.
The perceptual effects of reduced speech are not limited to intelligibility. A number of studies illustrate that reduced speech is processed more slowly or effortfully than unreduced speech. For example, both Ranbom and Connine (2007) and Tucker (2011) reported that reduced wordforms were responded to more slowly in a lexical decision task than unreduced wordforms. Reduced speech also appears to lead to lower lexical activation than unreduced speech: in cross-modal priming tasks, both Tucker and Warner (2007) and Ranbom et al. (2009) found that reduced forms are less effective primes than unreduced forms. Similarly, using a diverse array of methods, Pitt (2009) found that reduced forms lead to less lexical activation than citation (unreduced) forms.

Brouwer et al. (2012) theorized that these effects are in part due to increased lexical competition. They demonstrated that in a comparison of words like unreduced *computer* and reduced *puter*, the reduced form primed words like *pupil*, while the unreduced form did not. The greater number of potential competitors for the reduced form, they argued, leads to processing delay. While this explanation is plausible for reductions that involve wholesale deletion of phonemic content like *computer* ∼ *puter*, it is not clear the extent to which the explanation can account for other types of reductions, such as intervocalic spirantization or flapping, or gestural reductions that affect temporal coordination.

Sumner (2013) put forward the view that different processing strategies are used by listeners for different types of speech. Relatively clear, unreduced speech induces bottom-up processing which relies heavily on acoustic detail. Conversely, casual, reduced speech induces top-down processing which relies on contextual cues to recover meaning. This view is supported, in part, by the findings of Brouwer et al. (2013), who found that lexical access of reduced forms was facilitated by discourse context. The
observed effect size was proportional to the strength of the context. Crucially, citation (unreduced) forms were not affected by context, suggesting purely signal-driven processing.

It is also well-established that listeners have implicit knowledge of what reductions are possible or probable, and that this knowledge tempers their perceptual expectations: Mitterer and McQueen (2009) demonstrated that listeners’ implicit knowledge of the relationship between reduction and context affects their anticipatory eye movements; Mitterer and Russell (2013) found that listeners interpret reduction in real-time to make inferences about whether they are hearing a high- or low-frequency word; and LoCasto and Connine (2011) found that the size of priming effects is correlated with the probability of reduction of a particular form. However, the simple view of ‘reduction’ being a single dimension of variability may obscure a more complex state of affairs. Mitterer (2011) demonstrated that Dutch /t/-deletion and nasal place assimilation, two common processes of reduction, are perceptually processed by different mechanisms, the former being processed later than the latter. This finding is especially relevant since much of the literature on the perception of reduced wordforms has only examined American English word medial /t/-deletion/flapping, such as in words like center or water. The generalizability of these studies, therefore, to other processes of reduction may be questionable.

Taken together, this literature suggests that there are measurable and consistent effects of phonetic reduction on perception—compared to unreduced speech, reduced speech is processed slower and is less intelligible. However, with the notable exception of Lieberman’s (1963) study, the majority of these studies have examined the perception of artificially induced reductions. That is, the stimuli were recordings of a trained phonetician producing words in unreduced, citation form (e.g. with a fully released and aspirated [tʰ]) and words in reduced, casual form (e.g. with a flapped [ɾ]). Further,
the majority of the reductions examined are highly salient, such as /t/-flapping or syllable deletion. The type of reduction observed in the last word of (1) is often much more subtle, with the productions of the last words in (1) and (2) sometimes only tens of milliseconds different in overall word duration.

3.1.2 Individual differences in perception

It is well-known that individual variation exists in speech production (Ferguson, 2004, 2012; Ferguson and Quené, 2014; McCloy et al., 2015). As established in Chapter 2, some of this variation in production is systematically related to individual differences in cognitive traits such as theory of mind. Since our understanding of the links between individual differences and speech production is still in its infancy, it stands to reason that we may expect to observe similar relationships between individual differences and speech perception. Indeed, hints of such a relationship already exist within the literature (reviewed in the following sections), suggesting that it may be reasonable to expect that cognitive measures such as AQ and RMITE scores may be relevant for the perception of reduced speech.

3.1.2.1 Speech perception in adverse conditions

Perception of speech in noise has been shown to be more effortful than perception of speech in the clear. Murphy et al. (2000) found that word pairs were more difficult to recall when presented in babble relative to when they were presented in quiet, suggesting that the presence of noise required the use of cognitive resources that would otherwise have been employed in memory. Gilbert et al. (2014) additionally found that this memory deficit was smaller when participants were presented with speech

---

1 See Doherty et al. (2012) for a more general overview of the problem of unexplained variance in general within the psychological sciences.
which was originally produced in the presence of noise (i.e. clear speech) than when
the stimulus speech was originally produced in quiet. This finding illustrates an excit-
ing synergy between perception and production, whereby talker behavior goes some
way toward offsetting the listener’s disadvantages in adverse conditions. A conceptu-
ally similar finding was reported by Scarborough and Zellou (2013), who found that
response latencies in a lexical decision task were faster when responding to speech pro-
duced to a hard-of-hearing addressee than responses to speech produced to an imagined
hard-of-hearing addressee. This result speaks to the fact that although talkers may be
competent at optimizing their speech for a listener, their conscious control over these
processes is weak. This kind of link between production and perception is central to
the listener-oriented assumptions outlined above.

Memory load has been used as an independent variable in some studies of
speech processing. For example, Toro et al. (2005) presented participants with a stream
of artificial speech with statistical regularities which give cues to word boundaries (in
the style of Saffran et al., 1996). Some participants were told to passively listen to the
speech, while others performed some concurrent task. They found that the passive lis-
teners significantly outperformed the multitaskers in a subsequent word segmentation
task. These results suggest that when attentional resources are taxed, statistical learn-
ing is inhibited, leading to poorer word segmentation. This conclusion was elaborated
and made somewhat more complex by Mattys et al. (2009), who demonstrated that
attentional load should not be conceived of a single parameter. More particularly, cog-
nitive load, parameterized in this case as attention to a competing simultaneous task,
causes perceivers to lean more heavily on high-level, lexical cues in word segmenta-
tion, at the expense of low-level, acoustic cues. In the case of listening to phonetically
reduced speech or some other degraded speech signal, this result implies that, all else

2For a review of speech perception in adverse conditions in general, see Mattys et al. (2012).
being equal, greater cognitive load should lead to poorer perception. This poor perception may be reflected in recognition accuracy, response latency, judged clarity, or similar behavioral measures.

### 3.1.2.2 Speech perception under cognitive load

Early research in psychosyntax has established that individual differences in working memory ability influence sentence processing. Daneman and Carpenter (1983) found that working memory span was negatively correlated with the extent of slowdown when reading a sentence with a semantic oddity, such as *There is a sewer near our home who makes terrific suits.* King and Just (1991) similarly found that readers with shorter digit spans took longer to read sentences with a center-embedded relative clause (such as *The reporter that the senator attacked admitted the error*) than readers with longer digit spans.³ More recently, Haarmann et al. (2003) demonstrated that ‘conceptual span’, a measure of semantic working memory, correlates with performance in a wide variety of reading comprehension tasks.

Speech perception is also known to be affected by general cognitive functioning (see Akeroyd, 2008, for review). For example, accuracy in speech perception tasks was shown by Lunner (2003) to correlate with working memory span, and by Grant and Seitz (2000) to correlate with the ability to use contextual cues. A similar study by George et al. (2007) found that speech perception in noise correlates with text perception in noise (e.g. reading a sentence with barcode-like blocks superimposed), suggesting that recovering information from a noisy signal relies on general cognitive resources, regardless of whether the signal is auditory or visual. Much of the research in this domain has investigated the effects of the general cognitive decline experienced

---
³See also Lewis et al. (2006) and Sprouse et al. (2012) for more recent approaches to the problem of center-embeddings and memory load in sentence processing.
in adults of advanced age, rather than effects of deficits in specific cognitive abilities. In this line of work, Taler et al. (2010) found that older adults were more influenced by lexical factors than younger adults. That is, in a sentence repetition task, they responded more slowly and less accurately to sentences with words in dense phonological neighborhoods than to sentences with words in sparse phonological neighborhoods, and the size of this effect was correlated with age. This effect was interpreted as being due to the older adults’ decreased ability to inhibit lexical activation, presumably due to general cognitive decline. Other studies of aging (Lash et al., 2013) have claimed that a listener’s ability to use context declines with age (but cf. Dubno et al., 2000), which has similarly been explained as a decrease in inhibitory activity. This interpretation squares well with Engle and Kane’s (2004) proposal that executive control of attention is actually the underlying factor behind variation in both working memory and sentence comprehension; they claim that attentional control is the crucial element behind tasks that measure working memory span. Some empirical support for this perspective has been found in links between working memory, mind-wandering (as a measure of attentional control), and reading comprehension (McVay and Kane, 2012). This perspective is relevant to the present discussion of the perception of reduced speech insofar as we acknowledge that factors other than memory are relevant for individual differences in speech perception.

3.1.2.3 Interim summary

The research reviewed above comes from a variety of literatures: psychogerontology, cognitive psychology, audiology, and psychosyntax, but nevertheless it paints a coherent picture of speech perception. Adverse conditions, both external to the listener and internal, influence speech perception, and individual differences between listeners in cognitively-relevant dimensions such as working memory span also have a role to play.
Further, such influences affect speech perception at multiple levels of processing; the methodologies used in the studies reviewed above are diverse and target distinct levels of the speech chain.

In summary, according to listener-oriented theories, phonetic reduction is assumed to have universally negative effects on perceptual processes. In line with this claim, the empirical evidence generally supports the idea that reduced speech is harder to perceive and perceived more effortfully than unreduced speech. In Chapter 2, it was established that there are systematic individual differences in phonetic reduction which can be related to the cognitives measures of AQ and RMITE scores. Additionally, it has been established in a wide variety of literatures that variability in other cognitive measures, such as attentional load or working memory span, can influence speech perception and comprehension at a number of processing levels. It is therefore reasonable to predict that cognitive measures such as AQ and RMITE scores also influence the perception of reduced speech at multiple levels. An investigation of such factors could well hold insight for listener-oriented theories of reduction, such as that of Aylett and Turk (2004). As such, the present study aimed to assess the perception of reduced and unreduced speech produced by naïve talkers. Four distinct methodologies were used—subjective clarity judgements; speech intelligibility in noise; auditory lexical decision; and auditory semantic acceptability judgements—to investigate perception at various levels of speech comprehension. Further, this study sought to correlate individual differences in speech perception with variability in cognitive traits as measured by the AQ and RMITE scores.
3.2 Experiment 4: Speech intelligibility in noise and clarity judgements

3.2.1 Introduction

This experiment probed the assumption that reduced speech is less intelligible and less clear than unreduced speech. Speech intelligibility is classically measured via an identification in noise task, where a listener is presented with a stimulus masked by noise and is asked to identify the stimulus. Measures of subjective speech clarity are less well-studied, but one technique involves listeners judging clarity on a 7-point Likert scale (Eisenberg et al., 1998). Eisenberg et al. (1998) found that clarity judgements of word tokens are correlated with the tokens’ intelligibility. However, despite this correlation, these values differ in magnitude and therefore Eisenberg et al. (1998) cautioned against substituting one for the other. In another study, Ferguson and Kerr (2009) attempted to control for participants’ knowledge of the stimuli before the task began, thus making intelligibility a non-issue: since the participants knew the linguistic content of the stimuli, they should all be highly intelligible. Ferguson and Kerr (2009) reported that stimuli taken from ‘conversational’ or ‘everyday’ speech were rated as less clear than stimuli from ‘clear’ speech (addressed to a hearing-impaired listener), in spite of the similar intelligibility level. This finding suggests that subjective speech clarity judgements can capture elements of perception that are distinct from those of intelligibility (see also Preminger and Van Tasell, 1995).

This experiment therefore sought to determine whether unreduced speech, relative to reduced speech, is more intelligible and/or judged as being more clear. A secondary goal of this experiment was the investigation of the effects of individual differences in cognition on listener behavior.
3.2.2 Method

3.2.2.1 Participants

Thirty-five undergraduate students (28 female) from Ohio State University participated in the experiment for partial course credit. All participants were monolingual English speakers with no reported history of speech or hearing impairments, or autism spectrum conditions.

3.2.2.2 Stimulus materials

The acoustic stimulus materials in this experiment were drawn from the production data from Experiment 1a (described in Chapter 2). Naïve talkers produced nouns in sentence-final position either in a high-predictability (HP) or low-predictability (LP) context, such as the final words in (3) and (4) respectively. Eleven talkers were selected with either low AQ scores (from 90 to 104) or high AQ scores (118 to 128). Tokens from those talkers were extracted from the sentences and selected so as to maximize a $2 \times 2 \times 2$ design, crossing token duration (long or short) with talker AQ score (high or low) with sentence context predictability (high or low). Long and short duration tokens were taken from the fourth and first quartiles of the word duration distribution, respectively, calculated on a by-talker basis. Short stimuli ranged from 146ms to 289ms, and long stimuli ranged from 473ms to 790ms in duration. Although the selection strove to obtain as balanced a design as possible, maximizing the number of tokens was the primary priority; a total of 240 tokens were selected, and Table 3.1 provides the number of stimuli in each cell of the $2 \times 2 \times 2$ design.

(3) The house was robbed by a thief.

(4) The old woman discussed the thief.
Table 3.1: Number of stimuli in each cell of the $2 \times 2 \times 2$ design of the intelligibility and clarity experiment.

<table>
<thead>
<tr>
<th></th>
<th>High AQ talkers</th>
<th>Low AQ talkers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HP</td>
<td>LP</td>
</tr>
<tr>
<td>Long</td>
<td>39</td>
<td>24</td>
</tr>
<tr>
<td>Short</td>
<td>27</td>
<td>55</td>
</tr>
</tbody>
</table>

The three variables manipulated in the stimulus design—token duration, talker AQ score, and sentence context predictability—differ in their level of description. Token duration is a direct acoustic measure of the stimulus token; talker AQ score is a cognitive score of the talker producing the token; and sentence context predictability is a binary linguistic measure of the context the token was uttered in. Since the current experiment involves purely auditory stimuli out of context, the hypothetical listener does not know what the talker’s AQ score is or what sentence context the token was uttered in, unless, that is, these variables influence the acoustics of productions in such a way that they influence listener responses. Therefore, to the extent that the variables of talker AQ score and sentence context predictability influence listener responses, they can be interpreted as dimensions of *acoustic variability* which are unaccounted for by token duration.

After extraction from the sentences, each token was normalized to the same mean intensity. The same tokens were used for both the intelligibility task and the clarity task, however the tokens in the speech intelligibility task were mixed with speech-shaped noise at -2dB SNR; that is, the noise was 2dB louder than the signal. The spectrum of the noise was generated to match the LTAS of all stimuli averaged together; the amplitude of the noise did not vary. The tokens in the clarity judgement task were presented in the clear; the only signal processing performed on them was amplitude normalization.
3.2.2.3 Procedure

Each participant was seated at a computer with a keyboard and professional monitoring headphones in a quiet room with no more than four other participants. On each trial of the intelligibility task, the participant heard one word mixed with noise and was instructed to identify the word by typing it with the keyboard. The next trial was presented after the participant pressed the enter key—there was no time limit on individual trials.

Following the intelligibility task, the participants carried out the clarity rating task. They were instructed that they would hear words spoken without any background noise, and that they would rate how clearly each word was pronounced on a 1–7 scale. In order to decrease the possibility that low clarity ratings were simply artifacts of intelligibility issues (following Ferguson and Kerr, 2009), participants were shown the list of all of the stimulus words used in the clarity task, and instructed to read it carefully. This activity ensured that participants knew what the words they were hearing were supposed to be, and thus could make the subjective clarity judgement independently of concerns of intelligibility. The list of words was shown on-screen for 60 seconds before participants were able to advance; many participants studied the list for substantially longer than 60 seconds. On each trial, a single word was presented over headphones without noise, and participants responded by pressing one of the keys 1 through 7 on the keyboard. On screen, the scale was labeled following Ferguson and Kerr (2009): 1, minimum; 2, very unclear; 3, somewhat unclear; 4, midway; 5, clear; 6, very clear; 7, maximum.

Following these linguistic tasks, participants completed the AQ questionnaire and the RMITE test. All four tasks (intelligibility, clarity, AQ, RMITE) were self-paced
and would not advance to the next trial until a valid response had been given by the participant.

### 3.2.2.4 Analysis

The intelligibility responses were scored as correct or incorrect (no partial matches). Homophones (such as crews and cruise) were scored as correct. Incorrectly spelled words were marked as incorrect, due to the difficulty in objectively interpreting a participant’s intention from writing alone.

A mixed effects logistic regression model was constructed to predict intelligibility response accuracy. Fixed effects of talker AQ scores, word duration, predictability condition (HP vs. LP), and all two- and three-way interactions were entered. Additional fixed effects were listener AQ and eye-reading scores. The random effect structure consisted of random intercepts of listener and word, with random listener and word slopes for duration and predictability condition. All continuous variables were centered at zero before being entered into the model.

The responses in the clarity judgement were extremely non-normal, as shown in the histogram in Figure 3.1. Additionally, these data are clearly interval rather than continuous (i.e. a response of 5.5 is impossible), making a linear regression statistically inappropriate. The distribution in Figure 3.1 resembles a Poisson distribution flipped on its x-axis. Re-scaling the clarity scores so as to invert the x-axis (i.e. each response $x$ is scaled to $7 - x$) reveals a distribution that meets several characteristics of a Poisson distribution. As such, these data were modeled using a mixed effects Poisson regression to predict the integer response along the (rescaled) clarity scale (see Faraway, 2004).

---

4This pattern of results and high overall clarity ratings is consistent with the findings of Ferguson and Kerr (2009). Given that these responses were to words presented in the clear to adults with normal hearing, the reported clarity of the stimuli is not surprising.

5For example, the mean $\bar{x}$ and median $\tilde{x}$ satisfy the inequality $\tilde{x} - \frac{1}{3} < \bar{x} \leq \tilde{x} + \ln 2.$
Figure 3.1: Histogram of clarity judgement responses.

2006, chapter 3). The fixed and random effect structure was identical to that of the logistic model for intelligibility. As in the intelligibility model, all continuous variables were centered before being entered into the model.

Both the intelligibility and clarity models were selected using a stepdown procedure. Beginning with the fully-specified model, fixed effects with small $z$ values were removed from the model, their removal being justified (or denied) by likelihood ratio testing. This procedure continued until arriving at the smallest justified model. Simple effects were not removed if they were involved in significant interactions.
3.2.3 Results

3.2.3.1 Intelligibility

The model output for the mixed effects regression model of intelligibility is summarized in Table 3.2. The model revealed one significant simple effect and two significant interactions, all involving word duration. The significant simple effect of duration was that longer words, as expected, were more intelligible than shorter words. This finding is consistent with previous research which has found correlations between word duration and intelligibility (Bond and Moore, 1994; Hazan and Markham, 2004).

![Log likelihood: −3687](image)

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>−0.452</td>
<td>−1.742</td>
<td>.008</td>
</tr>
<tr>
<td>Talker AQ</td>
<td>0.001</td>
<td>0.280</td>
<td>.780</td>
</tr>
<tr>
<td>Duration</td>
<td>4.905</td>
<td>4.438</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Predictability: Low</td>
<td>−0.068</td>
<td>−0.379</td>
<td>.705</td>
</tr>
<tr>
<td>Talker AQ × Duration</td>
<td>−0.109</td>
<td>−7.044</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Duration × Predictability: Low</td>
<td>−2.326</td>
<td>−3.242</td>
<td>.001 **</td>
</tr>
</tbody>
</table>

Table 3.2: Model output for accuracy on intelligibility task.

A significant interaction between talker AQ score and duration was observed, such that words from talkers with a lower AQ score were more intelligible than words from talkers with a higher AQ score, but only for words with a longer duration. A visualization of this interaction is depicted in Figure 3.2, where it can be seen that the lack of an influence of talker AQ score for the short duration words is clearly a floor effect—overall intelligibility is close to zero. Although talker AQ score was entered into the regression model as a continuous variable, it is depicted as a discrete variable (high vs. low) in the figure for clarity.
The other significant interaction was between the word’s predictability condition and duration, such that words extracted from an HP context were more intelligible than words extracted from an LP context, but only for words with a longer duration. Figure 3.3 depicts this interaction. As with the talker AQ interaction mentioned above, it can be seen that the convergence of the HP and LP trend lines is plausibly due to a floor effect.

These two interaction effects suggest that words from talkers with lower AQ scores are more intelligible than words from talkers with higher AQ scores, and that words extracted from HP contexts are more intelligible than words extracted from LPcontexts.
Figure 3.3: Word intelligibility as a function of duration, split by talker word production condition. Curves are best fit binomials, grey bar is bootstrapped 95% confidence interval.

contexts. The context effect is somewhat unexpected—given that words in HP contexts tend to be reduced, one might reasonably expect the HP words to be less intelligible than the LP words. However, the modeling procedure here controlled for word duration, a measure of reduction. These effects must therefore reflect some aspect of the acoustic signal, such as spectral properties or within-segment temporal detail, which systematically varies between both individual talkers and predictability contexts, such that tokens from talkers with low AQ scores are more intelligible than tokens from talkers with high AQ scores, and that, counterintuitively, tokens from the HP context are more intelligible than tokens from the LP context.
3.2.3.2 Clarity

Log likelihood: $-5140$

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$z$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>$-0.261$</td>
<td>$-2.014$</td>
<td>.044</td>
</tr>
<tr>
<td>Talker AQ</td>
<td>$0.007$</td>
<td>$6.366$</td>
<td>$&lt;.001$</td>
</tr>
<tr>
<td>Duration</td>
<td>$-3.465$</td>
<td>$-7.129$</td>
<td>$&lt;.001$</td>
</tr>
<tr>
<td>Predictability: LP</td>
<td>$-0.022$</td>
<td>$-0.337$</td>
<td>.736</td>
</tr>
<tr>
<td>Talker AQ $\times$ Duration</td>
<td>$0.043$</td>
<td>$7.634$</td>
<td>$&lt;.001$</td>
</tr>
<tr>
<td>Duration $\times$ Predictability: Low</td>
<td>$1.214$</td>
<td>$4.499$</td>
<td>$&lt;.001$</td>
</tr>
</tbody>
</table>

Table 3.3: Model output for clarity rating task.

The results for the clarity task were similar to those of the intelligibility task. The model output is summarized in Table 3.3; recall when interpreting the coefficients that the ratings were re-scaled for analysis, such that a negative coefficient represents a clearer rating. All graphs, however, preserve the original scale, where 7 represents the most clear and 1 the least clear. A significant simple effect of duration was observed, such that longer words were rated as clearer than shorter words. A significant simple effect of talker AQ score was also observed, such that words from talkers with higher AQ scores were rated as less clear than words from talkers with lower AQ scores. However, both of these effects can only be interpreted in light of their interaction, which was also significant. These two factors, and their effect on perceived clarity, are depicted in Figure 3.4. As can be seen in the figure, the clarity benefit for words from low AQ talkers only holds for words longer than approximately 400ms. For the shorter words, there is no difference in clarity ratings between the words from the different talkers.

A significant interaction between duration and word predictability condition was also observed, such that words excised from an HP context were judged clearer
Figure 3.4: Perceived word clarity as a function of duration, split by talker AQ score. Curves are best fit Poisson regression, grey bar is bootstrapped 95% confidence interval.

than words from an LP context, but only for words of a relatively long duration. This interaction is depicted in Figure 3.5. Note that this interaction is very similar to the effects of these variables on intelligibility, depicted in Figure 3.3, whereby long duration words from an HP context were more intelligible than long duration words from an LP context.

3.2.4 Discussion

This experiment revealed expected effects of duration on both intelligibility and clarity judgements: words with longer durations were more intelligible and were judged to
Figure 3.5: Perceived word clarity as a function of duration, split by talker word production condition. Curves are best fit Poisson regression, grey bar is bootstrapped 95% confidence interval.

be more clear than words with shorter durations. These results are consistent with those of Lieberman (1963), and are consistent with the assumption of listener-oriented theories that unreduced speech is perceptually beneficial relative to reduced speech.

Additionally, words from talkers with low AQ scores were found to be somewhat more intelligible and rated as clearer than words from talkers with high AQ scores. This effect of talker AQ score on both intelligibility and clarity suggests that the lower AQ talkers, whose speech was more intelligible and judged clearer, have generally more spectral clarity than the higher AQ talkers. This conclusion is motivated by the fact that duration was controlled for in this task, which means that the
observed effect must be due to some other (set of) acoustic parameter(s). Experiment 1a in Chapter 2, where these stimuli originate from, did not find a simple effect of AQ on vowel dispersion—that is, it is not the case that talkers with lower AQ scores have generally more disperse vowels than higher AQ talkers, suggesting that the spectral characteristics leading to the intelligibility benefit are not related to vowel dispersion. An alternative explanation for this effect could be that lower AQ talkers are just ‘better at talking’, since they are generally more social than high AQ talkers. However, such an interpretation cannot be reduced to one of ‘experience with talking’, otherwise we would expect speech intelligibility to be a monotonic function of talker age, which it is not (Hazan and Markham, 2004; Holliday et al., 2015; Romeo et al., 2013). Additionally, the mechanisms linking sociability to language production are unclear (Dewaele and Furnham, 2000; Slomkowski et al., 1992). As such, any conclusion drawn from this effect must be treated with great caution.

A similar effect was observed in which words from HP contexts were more intelligible and rated as clearer than words from LP contexts. This effect is unusual from a perspective which views HP words as reduced and LP words as unreduced. However, as mentioned in the results section, since word duration was controlled for, it is not necessarily true that the HP/LP distinction in these stimuli is really a measure of reduction, especially if reduction due to semantic predictability is restricted primarily to temporal effects (e.g. Watson, 2010).

Considering the selection of the stimuli in the experiment design may assist in approaching an explanation for this effect. The stimuli were selected for this experiment based on their durations; only words in the top 25% and bottom 25% of the duration distribution were chosen. Additionally, in the dataset that the words were drawn from, LP words were in general longer than HP words. Therefore, all else being equal, a word of (say) 400ms duration was likely to be relatively long for an HP word and
relatively short for an LP word. If we assume that spectral clarity is a function of temporal typicality (Moon and Lindblom, 1994), then it follows that a longer-than-average HP word is expected to be relatively spectrally informative, and that a shorter-than-average LP word is expected to be relatively spectrally uninformative, even if both of these words have the same overall duration (and, indeed, even if both words are the same lexical item). This difference in spectral clarity can then explain the observed intelligibility and clarity benefit for the HP words.

However, such an explanation relies on the production mechanisms for LP and HP words being distinct and discrete, rather than continuous functions. It also relies on the noted assumption that spectral clarity is a function of temporal typicality: while this assumption may seem reasonable, naïve assumptions about how speech production mechanisms relate to each other often prove false (see e.g. Scarborough, 2013). Indeed, these results could be interpreted to suggest that temporal and spectral reduction are not be related at all. If the LP and HP words differed significantly in their spectral informativity, a simple effect of condition (HP < LP) would be expected, after controlling for duration. However, only interactions were observed, suggesting that the type of reduction observed in (1) is purely temporal, or that listeners are not sensitive to these spectral cues.

Finally, the lack of any observed effects of theory of mind or autistic traits of the listeners in the present data suggest that autistic traits play a minimal role in speech intelligibility. This conclusion may appear contrary to Alcántara et al.’s (2004) finding that individuals with high functioning autism tended to have slightly poorer speech perception in noise than neurotypical controls. However, these experiments are not directly comparable: all of the participants in the present study were neurotypical, i.e. Alcántara et al.’s (2004) control group. Neurotypical individuals with high AQ scores
cannot be compared to individuals with high-functioning autism without significant independent evidence.

3.3 Experiment 5: Lexical decision

3.3.1 Introduction

This experiment addressed the question of whether unreduced words confer a speed or accuracy benefit in auditory lexical decision relative to reduced words. Ranbom and Connine (2007) found that unreduced word forms are responded to more quickly and accurately than reduced forms. However, their stimulus materials involved relatively unsubtle reductions such as /t/-flapping, which raises the question of whether such effects can be observed for more subtle reductions such as the ones involved in the stimuli for experiment 4. Additionally, this experiment probed the role of individual variation in AQ and RMITE scores in modulating the listeners’ responses.

3.3.2 Method

3.3.2.1 Participants

Thirty-three undergraduate students (24 female) from Ohio State University participated in the experiment for partial course credit. All participants were monolingual English speakers with no reported history of speech or hearing impairments. None of the participants had participated in experiment 4.
3.3.2.2 Stimulus materials

Target words for this experiment were a subset of the target words used in experiment 4. A total of 122 tokens were selected, consisting of 46 unique lexical items from 11 different talkers. The tokens were selected to ensure a relatively even balance of predictability context (HP or LP), talker AQ score, and word duration, to the extent that such a balance was possible with these materials. These 122 tokens constituted the target words for the lexical decision task.

Since the source recordings consisted entirely of lexical words, nonwords had to be ‘constructed’ from the recordings. For example, productions of /Jim/ and /tivz/ were extracted from the phrases ‘she might’ and ‘detectives’ respectively. Care was taken to ensure that the duration distribution of the nonwords was similar to that of the lexical words. A total of 113 tokens were extracted for use as nonwords. Fifteen of these nonwords involved the vowel /ɔ/, for example /gɔr/ (extracted from burglar). However, none of the target words in the source materials had this vowel in them. To avoid this vowel being a signal for nonwordiness, four filler word tokens with /ɔ/ were added, which were productions of bird, nurse (×2), and work.

There were up to three repetitions of the same lexical item in this set of materials. These repetitions were necessitated by the relatively small size of the source materials. In order to diminish potential repetition effects on participant responses, the items were split into three blocks of 80, 80, and 79 items each. Each lexical item was presented only once per block. In addition, each block was pseudorandomized such that there were no more than 3 consecutive word or nonword tokens, and no more than 2 consecutive stimuli from the same talker. These three blocks were counterbalanced by Latin square, for three orders of presentation. Additionally, half of the participants were presented with each block in reverse order, for a total of six distinct
subject lists (3 block orders × 2 block-internal orders). Each list had five participants, with the exception of one which had eight.

### 3.3.2.3 Procedure

On each trial, the stimulus was presented over headphones at a comfortable volume. Participants were instructed to answer with a ‘word’ or ‘nonword’ response via a button box as quickly and as accurately as possible. The task was self-paced, and the experiment would not continue until the box had registered a response from the participant. Between the first and second blocks, the listeners completed the AQ questionnaire, and between the second and third blocks, the RMITE test. These tasks, as well as providing important information on individual differences, also allowed the participants to take a mental break from the lexical decision. Reaction time (RT) and response accuracy were both measured.

### 3.3.2.4 Analysis

Accuracy rates were analyzed via logistic mixed effects regression models on target (i.e. word) trials only. Fixed effects reflected properties of the stimulus and properties of the listener. The stimulus properties were intelligibility (taken from experiment 4 above), repetition (1st, 2nd, or 3rd presentation of this lexical item), stimulus production context (HP or LP), stimulus word duration, and talker AQ score; the listener properties were listener AQ score and listener RMITE score. In addition to those fixed effects, all two-way interactions within the stimulus properties were added to the model, and also two- and three-way interactions between the listener properties and the stimulus properties and their interactions. In R-style Wilkinson notation, the fixed effect structure
can be written as follows:

\[(\text{Intelligibility} + \text{Repetition} + \text{Production Context} + \text{Duration} + \text{Talker AQ.c})^2 \times (\text{AQ Score} + \text{RMITE Score})\]

All variables were centered around the mean before being entered into the model. Additionally, random intercepts of listener, talker, and word identity were added. Owing to the complexity of the model, random slopes were not included. The same stepdown procedure used in the analysis of experiment 4 was used.

Preliminary modeling revealed a spurious interaction between stimulus production context and listener AQ score. Figure 3.6 shows correct response proportion as a function of listener AQ score, split by stimulus production context. Note that the HP trend line has a much more negative slope than the LP trend line, suggesting that high-AQ listeners were more inaccurate for HP words than were the low-AQ listeners. However, this trend is driven almost entirely by a sole outlier in the 130+AQ range, who performed very poorly with the HP words but relatively well with the LP words. This spurious result therefore motivates the inclusion in the model of a random slope for predictability by listener, and this model is the one reported in the results section.

Reaction times were also measured. Because variability in word duration was key to the experimental design, it was therefore of paramount importance to ensure that reaction time measures were interpretable. Reaction time was measured from word onset, and because the durations of each token were known, it was therefore possible to also compute RTs from word offset and word midpoint. However, even with this adjustment it is difficult to compare RTs from words of different durations, since words with longer duration are necessarily accompanied by more time for the listener to parse and process the input. It is worth considering in some detail the possible interpretations that come with measuring reaction time in these different locations.
Figure 3.6: Proportion correct response in the lexical decision task as a function of listener AQ score, split by stimulus production context. Note the outlier with AQ > 130 whose poor accuracy with HP targets shifts the overall slope of the trend.

The predicted differences between these three measures of RT with respect to stimulus duration depends on one’s theory of processing. We will proceed here by creating models to predict RT from word offset (and then calculate mid and onset RTs from the offset RT). This choice of measure to be ‘primary’ is motivated by the assumption that the response must be after the word offset. This assumption is reasonable in the case of this experiment: first, all the stimulus words are monosyllables where the uniqueness point and word offset are the same point in time; second, in the context of a lexical decision task, stimuli include nonwords, and uniqueness points are therefore unreliable in determining lexical identity of items. Given these facts, responses prior to word offset are unexpected, since the participant cannot truly be sure that the stimulus is indeed a word.
First, we consider the null hypothesis—that all words are processed equally quickly. This model can be formalized by stating that the predicted RT is simply some positive constant $c$, that is:

$$RT_{offset} = c$$

A more plausible model is one where word duration influences reaction time. The most straightforward approach is a linear model, where some coefficient $b$ linearly applies to the influence of $d$ on $RT$, i.e.:

$$RT_{offset} = b \cdot d + c$$

It can be seen that the ‘constant’ model above is simply a special case of the linear model where $b = 0$. Figure 3.7 shows predicted RT values for all three of our measuring points for four different values of $b$, keeping $c$ constant.\(^6\) The differing values of $b$ reflect assumptions about the role of signal duration in lexical processing. Under the constant model ($b = 0$), word duration does not influence reaction time from word offset. However, from measures earlier in the word (onset and mid), a linear trend can be observed. This model represents the theory that all words, regardless of acoustic content, are processed equally quickly and that processing does not begin until the word offset. Here, measuring RT from earlier in the word leads to a positive slope in the graphs, since processing does not begin until the word offset.

Predicted RTs in a model with a positive $b$ value ($b = 1$) are shown in the top left of Figure 3.7. This model reflects a theory whereby words with longer duration (more physical signal) take longer to process; that is, the extra signal causes some kind of slowdown. The processing system takes longer to wade through more information than it does through less. All three RT measures are monotonically increasing in this case.

\(^6\)The onset RTs are calculated as $RT_{offset} + d$ and the mid RTs as $RT_{offset} + d/2$. 

146
Figure 3.7: Predicted reaction time as a function of word duration under four different linear models.

Models with negative $b$ values are shown in the bottom left ($b = -0.5$) and bottom right ($b = -1$) of Figure 3.7. These models reflect the theory that shorter words take longer to process, or in other words, that longer words provide more time to begin processing before the word offset, thus allowing for a speedier post-offset response. This kind of model is most concordant with contemporary theories of lexical processing and cognition (e.g. the cohort model Marslen-Wilson and Tyler, 1980). It can be seen that the choice of $b$ will affect the slope of the onset and mid RTs, such
that positive slopes are possible. For all negative $b$ values, however, the offset RT is guaranteed to be negative. Therefore, it can be seen that it is possible for these RT measures to show different effects of duration—the onset RT can have a positive slope, the mid RT can have a zero slope, and the offset RT can have a negative slope, as when $b = -0.5$—and the results still be consistent with a simple model of lexical processing. We will return to these models in the results section.

Log-transformed reaction times, calculated from word offset and log transformed, were analyzed via linear mixed effects regression modeling. Only correctly-answered target (word) trials were included in the analysis. The model structure was identical to that of the logistic model for accuracy described above: all fixed effects for stimulus properties and their two-way interactions, all listener properties, and the two- and three-way interactions between the listener properties and the stimulus properties; random intercepts for listener, talker, and word. The model selection procedure was likewise identical. All variables were centered around the mean before being entered into the model. Markov chain Monte Carlo sampling was used to obtain $p$-values (Baayen, 2008).

Prior to analysis, all tokens with lower than chance accuracy (i.e. <50%) were removed, resulting in the removal of 11 target items (8.6% of total). Additionally, one participant who scored less than 70% accuracy was removed from the analysis. A total of 3,569 data points were therefore included in the accuracy analysis. To remove extreme outliers, all offset RTs more than 3 standard deviations away from each participant's mean, and all offset RTs less than 100ms, were removed, resulting in the removal of RTs from 252 trials, leaving a total of 2,890 trials with RTs available for analysis. The removal of an offset RT also meant the removal of the mid and onset RTs for that trial.
3.3.3 Results

3.3.3.1 Accuracy

Overall accuracy, for all stimuli, was 87.2%; accuracy on target trials was 88.0%. The final, simplest model for the accuracy data consisted of fixed effects of word repetition, duration, and intelligibility, with no significant interactions. The model output is shown in Table 3.4. This model (log likelihood: $-1079.1$) did not differ significantly from the fully-specified model described in the analysis section above (log likelihood: $-1050.8$; $\chi^2(44) = 56.574, p = 0.097$).

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$z$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.523</td>
<td>9.858</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Word repetition</td>
<td>0.207</td>
<td>2.898</td>
<td>.004**</td>
</tr>
<tr>
<td>Word duration</td>
<td>0.004</td>
<td>2.933</td>
<td>.003**</td>
</tr>
<tr>
<td>Word intelligibility</td>
<td>1.505</td>
<td>3.104</td>
<td>.002**</td>
</tr>
</tbody>
</table>

Table 3.4: Model output for lexical decision accuracy.

As mentioned in the stimulus section above, the target stimuli consisted of 122 tokens of 46 unique lexical items. This distribution necessitated the repetition of some lexical items, although the repeated items were always presented in different blocks. The regression model revealed that over the course of the experiment, accuracy improved. An accuracy boost of approximately 1% per repetition was observed: 86.7% for the first presentation, 87.7% for the second presentation, and 89.8% for the third presentation. It is not clear whether this effect is due to repetition of lexical items (McLennan and Luce, 2005), or simply a practice effect. Neither interpretation impacts the interpretation of the other results.
The effects of duration and intelligibility were such that words with longer duration, and words with higher intelligibility, were responded to more accurately than words with shorter duration and words with lower intelligibility, respectively. These effects are visualized in Figure 3.8 for duration and Figure 3.9 for intelligibility, and are consistent with the hypothesis that unreduced speech (words with longer durations and higher intelligibility scores) is beneficial to lexical access.

### 3.3.3.2 Reaction time

The output for the final, simplest model for the reaction time data is shown in Table 3.5. Recall that RT was log-transformed and centered prior to analysis, hence the small values of all the coefficients; all of the graphs in this section display RT on a log scale. This model (log likelihood: $-1721.1$) did not differ significantly from the fully-specified model described in the analysis section above (log likelihood: $-1699.2$; $\chi^2(37) = 43.787, p = 0.206$).

![Figure 3.8: Lexical decision accuracy on target words as a function of word duration.](image)
Figure 3.9: Lexical decision accuracy on target words as a function of word intelligibility.

Three simple effects and four interactions were observed for the RT data. As expected, more intelligible words were responded to faster than less intelligible words; see Figure 3.10. Expected simple effects of word repetition and duration were also observed: longer words were responded to faster than shorter words, and responses were overall faster later in the experiment relative to earlier in the experiment. Both of these effects are graphed in Figure 3.11, where the interaction between these factors can also be observed: the facilitation due to repetition was larger for longer words than it was for shorter words. All of these effects are consistent with the accuracy analysis, and with the notion that unreduced speech facilitates lexical access. Indeed, the interaction between duration and repetition suggests that unreduced speech enjoys more detailed memory representations and allows faster or more effective access to memory, consistent with Murphy et al.'s (2000) finding that speech perceived in the clear is more well-remembered than speech perceived in noise.
Effect

\( \beta \) & \( t \) & \( P_{MCMC} \) \\
(Intercept) & \( 9.868 \times 10^{-2} \) & 2.209 & .034 * \\
Intelligibility & \( -1.431 \times 10^{-1} \) & -2.601 & .011 * \\
Repetition & \( -8.759 \times 10^{-2} \) & -8.605 & < .001 *** \\
Duration & \( -1.809 \times 10^{-3} \) & -12.283 & < .001 *** \\
Predictability: LP & \( -2.591 \times 10^{-3} \) & -0.127 & .974 \\
Talker AQ & \( 3.552 \times 10^{-3} \) & 2.304 & .050 \\
Listener AQ & \( 1.005 \times 10^{-3} \) & 0.347 & .716 \\
Repetition \times Duration & \( -2.127 \times 10^{-4} \) & -3.430 & < .001 *** \\
Predictability: LP \times Talker AQ & \( -2.946 \times 10^{-3} \) & -2.163 & .030 * \\
Repetition \times Talker AQ & \( -1.926 \times 10^{-3} \) & -2.672 & .008 ** \\
Duration \times Listener AQ & \( -9.908 \times 10^{-6} \) & -2.484 & .016 * \\

Table 3.5: Model output for lexical decision reaction times from word offset.

Figure 3.10: Lexical decision reaction times to correct responses to target words as a function of word intelligibility.
Figure 3.11: Lexical decision reaction times to correct responses to target words as a function of word duration, split by block.

Two of the significant interactions involved talker AQ score. There is no known acoustic correlate of AQ score, nor is there any reason to suspect that one exists. Therefore, the variable of talker AQ score is best interpreted as a ‘catch-all’ category, accounting for acoustic variance in the stimulus which is not captured by the direct acoustic measure of duration, the indirect perceptual measure of intelligibility, or the indirect contextual measure of production context. For ‘talker AQ’ the reader is invited to substitute ‘miscellaneous acoustics systematically related to talker AQ score’. Larger claims and sweeping generalizations from this variable are discouraged, however, owing to the fact that the present data, although involving several thousand responses from listeners, only comprise 122 stimulus tokens from 11 talkers. With these caveats in mind, we will now turn to the interaction between word repetition and talker AQ, visualized
in Figure 3.12. This effect is small but significant: greater repetition effects were observed for responses to stimuli from high-AQ takers than for responses to stimuli from low-AQ talkers. This result is difficult to interpret not only because of the issues relating to talker AQ discussed above, but because of the experimental design. Repeated lexical items were not necessarily from the same talker; indeed, in the majority of cases, repeated lexical items were from different talkers. Given the small number of talkers in the stimulus set, it is difficult to justify drawing a conclusion about high-AQ talkers versus low-AQ talkers, but rather we are forced to conclude that some talker characteristics in the stimuli facilitated reaction time due to word repetition, while other talker characteristics did not confer the same advantage.

The other interaction involving talker AQ was with word production context (HP or LP), and is visualized in Figure 3.13. This effect is small but significant, and
suggests that the distinction between HP and LP words leads to reaction time differences only in stimuli from talkers with high AQ scores. Specifically, when the words were spoken by a high-AQ talker, LP words were responded to faster than HP words. When the words were spoken by a low-AQ talker, there was no reaction time difference between the word types. Since, as discussed above, talker AQ is an analogue for spectral differences not covered by other acoustic measures, this graph reaffirms the findings of Chapter 2. Namely, this result suggests that high-AQ talkers produce more differences between the HP and LP contexts than do the low-AQ talkers, and that these acoustic differences are indeed perceptually relevant.

The final significant interaction observed in the model was that between duration and listener AQ. Listeners with higher AQ scores had a larger influence of duration on their RTs than listeners with lower AQ scores. That is, longer words were responded
to faster, and the size of this effect was related to a listener's AQ score. This effect is visualized in Figure 3.14, which shows RT as a function of listener AQ, split by word duration (converted to a binary measure of long (> 400ms) versus short (< 400ms)). It can be seen that the listeners with higher AQ scores experienced a larger facilitation effect of duration than listeners with lower AQ scores—that is, the distance between the trend lines is larger for high-AQ listeners than it is for low-AQ listeners.

We now return to the discussion of different methods of measuring reaction time. A simple linear model predicting offset RT from duration revealed values of $b = -0.69$ and $c = 733$. The predictions from that model for the other two RT measures are displayed in the left panel of Figure 3.15, and the observed data are displayed in the right panel. It can be seen that this very naïve model performs well at predicting
Figure 3.15: Lexical decision reaction times to correct responses to target words as a function of word duration. Left panel shows simulated data with model parameters $b = -0.69$, $c = 733$; right panel shows linear trend lines for the observed data.

the slopes of the midpoint RT and onset RT. These results overall confirm the finding reported in the regression model, that unreduced speech does indeed confer a perceptual benefit relative to reduced speech.

3.3.4 Discussion

As predicted under the assumptions of the listener-oriented account, unreduced speech facilitates lexical access relative to reduced speech. Words which were longer in duration or higher in intelligibility were responded to more accurately and more quickly than words which were shorter in duration or lower in intelligibility. These effects on reaction time cannot be attributed to the extra processing time afforded by longer
words. As Figure 3.15 shows, every 10ms of extra duration speeds processing by approximately 7ms. Put another way, the addition of 10ms of signal only adds an extra 3ms to response latency.

Individual differences between the listeners also led to some peculiar effects on reaction time. Listeners with higher AQ scores exhibited a larger RT facilitation effect from duration than listeners with lower AQ scores, as seen in Figure 3.14. From this figure it can be seen that this effect is best interpreted as high-AQ listeners undergoing a slowdown for reduced words relative to low-AQ listeners. This result is best interpreted in light of the wider autism phenotype; in particular, we focus on three empirical results. First, there is evidence that people with high-functioning autism are less good at dealing with speech in noise than are neurotypical controls (Alcántara et al., 2004). The results of the experiments in this chapter have so far confirmed the notion that reduced speech is similar to speech in noise, in that it is harder to perceive than unreduced speech. Under this view, reduction is itself a kind of ‘noise’. Additionally, people with high-functioning autism tend to pay relatively more attention to the acoustic content of speech than the linguistic content of speech, compared to neurotypical controls (Järvinen-Pasley et al., 2008). Within the context of the current task, participants with higher AQ scores may have been more apt to attend to the individual acoustic features of the stimulus tokens, rather than the (non)lexical content of the tokens, missing ‘the wood for the trees’ as it were. Finally, Stewart and Ota (2008) have demonstrated that people with higher AQ scores are less likely to be influenced by lexical knowledge in their phonetic perception, suggesting a weaker or less effective integration of acoustic and lexical cues, consistent with the general findings in the literature of non-holistic processing in people with autism (Frith and Happé, 1994). Taken together, these results suggest that people with high AQ scores, when listening to reduced speech, may have less ability to use contextual cues, such as knowledge from the lexicon on what
phonological structures are likely, or what words are frequent. This deficit manifests itself in the RT slowdown.

### 3.4 Experiment 6: Sentence acceptability judgements

#### 3.4.1 Introduction

Experiment 4 established that unreduced words are more intelligible than reduced words, and experiment 5 extended this result to lexical access, demonstrating that unreduced words are accessed more quickly than reduced words. This experiment sought to determine whether the benefits of unreduced speech are also observed in a semantic judgement task. Here, participants were presented with an auditory sentence and were asked to judge whether or not the sentence ‘made sense’. This task can be considered a syntacto-semantic analogue to the lexical decision task, where the stimuli were sentence-length, rather than word-length, and participants judged whether the stimulus is a sentence, rather than a word.

To the extent that sentence processing also involves the processes of lexical access and retrieval that are inherent to the lexical decision task, we expect to see facilitatory effects of unreduced speech. However, by embedding these (un)reduced words in a sentence context, the linguistic context and processing demands are substantially altered. Classic research in psycholinguistics emphasizes this point. For example, Shields et al. (1974) documented a phoneme monitoring task where the target phoneme was embedded in both pitch accented and unaccented syllables. The target syllables were in turn embedded in a larger context which was either a grammatical sentence or a string of nonsense words. Listeners responded faster to phonemes embedded in accented syllables than in unaccented syllables, but only when the global context was sentential.
In the nonsense word condition, accent conferred no benefit. Cutler and Foss (1977) further demonstrated that accenting influences both content and function words, and concluded that this benefit derives directly from the fact that listeners are able to use semantic and syntactic cues to predict upcoming material. These studies suggest that conclusions drawn from single-word presentation, such as the auditory lexical decision task in experiment 5, do not necessarily neatly generalize to sentence contexts.

This experiment sought to determine if phonetic reduction has any influence on the speed or accuracy of semantic acceptability judgements of sentences. Additionally, this experiment investigated the role of individual differences in AQ and RMITE scores in modulating listener responses.

3.4.2 Method

3.4.2.1 Participants

Thirty-nine undergraduate students (29 female) from Ohio State University participated in the experiment for partial course credit. All participants were monolingual English speakers with no reported history of speech or hearing impairments. None of the participants had participated in either of the two previous experiments.

3.4.2.2 Stimulus materials

To test the effects of reduction on sentence acceptability in a controlled manner, ideal stimuli should involve sentences with words of varying degrees of reduction. The stimuli for the present study were taken from the same source as those of experiments 4 and 5: experiment 1a in Chapter 2. The original recordings consisted of full sentences such as (5) and (6), where we expect reduction in the word cake in (6) but not (5).
(5) Tom wants to know about the cake.

(6) For your birthday I baked a cake.

Simply using the original sentences as recorded unfortunately conflates the reduction of the final word with the semantic content of the sentence. To control for potential confounds, stimuli were created by cross-splicing. Each sentence pair has two sentence contexts, as in (7) and (8).

(7) {Tom wants to know about the}

(8) {For your birthday I baked a}

Although these contexts can be referred to as ‘low’ versus ‘high’ predictability, respectively, to avoid confusion with the original context of the final word, sentence contexts will be referred to as being ‘neutral’, as (7), or ‘biased’, as (8). This terminology can be thought of as referring to the expectations of the listener upon hearing the sentence. In addition to the contexts, each sentence also provides a final word, either produced in a low-predictability context, like (5), or a high-predictability context, like (6). Thus, there are four types of cross-spliced stimuli possible: neutral–LP; neutral–HP; biased–LP; and biased–HP.

(9) Tom wants to know about the \{\text{cake}_{LP}\} \quad \text{(neutral–LP)}

(10) Tom wants to know about the \{\text{cake}_{HP}\} \quad \text{(neutral–HP)}

(11) For your birthday I baked a \{\text{cake}_{LP}\} \quad \text{(biased–LP)}

(12) For your birthday I baked a \{\text{cake}_{HP}\} \quad \text{(biased–HP)}

A total of thirty pairs of sentences were chosen from 10 talkers from the original recordings. Sentence pairs were chosen so as to maximize acoustic differences between the
productions of the final word. These stimuli were spliced at the zero-crossing of the final word onset boundary, to create 120 stimuli tokens. Twenty-eight ungrammatical sentences, to be used as fillers, were created by splicing semantically or syntactically incongruous words into the beginning, middle, or end of the sentence. Care was taken to ensure that the spliced items had similar initial and final phones to the items they replaced, to guard against incongruous formant transitions. In the following examples, the original word is indicated to the left of the slash, while the new spliced word is indicated to the right:

(13) {The man / The ship} spoke about the clue.

(14) Mary hasn’t {discussed / slept} a blade.

(15) I made the phone call from a {booth / pizza}.

With this setup, the experiment was vulnerable to the confound that all non-spliced items are grammatical; thus, if a participant is able to detect splicing, their responses will be acoustically motivated, rather than syntactically or semantically motivated. One approach to this problem could have been to record some new talkers producing ungrammatical, non-spliced utterances; but such an approach would lead to participants developing talker-specific response strategies—it is known that naïve listeners are good at detecting and identifying different talkers even in degraded conditions (e.g. Remez et al., 1997). As such, the best approach to dealing with this problem was to include in the experiment a manipulation to test whether or not the listeners are indeed able to detect spliced speech. All of the low-predictability sentences were constructed such that any noun can end the sentence in a congruous manner, see e.g. (7) above. This construction allowed the splicing of any word to the end of one of these sentences

---

7 All non-spliced items are grammatical, but not all grammatical items are non-spliced. That is, some grammatical items were spliced, as described.
(modulo the formant transitions on the offset of the word the, see Stickney and Assmann, 2000) and the sentence is both grammatical and congruous. As such, responses to the unspliced original sentences (e.g. Tom wants to know about the cake$_{LP}$) can be compared with responses to the spliced but congruent sentences (e.g. Tom wants to know about the cake$_{HP}$). If the participants are sensitive to the spliced speech, we expect slower reaction times for the spliced stimuli than for the non-spliced original stimuli (Fowler, 1984; White et al., 2014). If the participants cannot detect the spliced speech, then their reaction times to both sets will be the same.

Stimulus tokens were balanced into four lists for a between-subjects design, so that each participant only heard one sentence with a given final word, i.e. 30 target tokens. The lists were pseudo-randomized so that no more than three target or filler trials were sequential, and there were no more than two sequential trials with the same talker voice.

3.4.2.3 Procedure

Each participant was seated at a computer with a button box and professional monitoring headphones in a quiet room with no more than four other participants. Participants were instructed that they would hear sentences over the headphones, and had to respond to the question ‘does this sentence make sense’ by pressing ‘yes’ or ‘no’ on the button box as quickly and accurately as possible. At the beginning of the experiment, listeners had four practice trials with feedback. The purpose of these trials was to give the participants an idea of the kinds of sentences they would hear, and what counts as ‘making sense’ for the purposes of the experiment. The practice sentences were as follows:

- “Household goods are moved in a flame.” (correct answer: no)
- “David does not discuss the hug.” (correct answer: yes)
• “They fished knew about the lid.” (correct answer: no)
• “Stir your coffee with a spoon.” (correct answer: yes)

After the practice, the test trials began. The test trials did not have feedback, and were self-paced with no time limit. Reaction time from sentence offset and response accuracy were recorded. After completing all 57 trials (30 target + 27 filler), participants completed the AQ and RMITE tasks.

3.4.2.4 Analysis

Accuracy on target trials was modeled via logistic mixed effects regression modeling. Fixed effects consisted of the stimulus properties of final word duration, final word production context (HP or LP), sentence context (biased or neutral), and talker AQ score, and the listener properties of listener AQ score and RMITE score. Two-way interactions amongst the stimulus properties were included, as were two- and three-way interactions between the listener properties and the stimulus properties. This fixed effect structure is the same as that of experiment 5, in terms of the interactions between and among listener and stimulus properties. Random intercepts of listener and word identity were included. All continuous variables were centered around the mean before being entered into the model. The model selection routine was the same as that followed for experiment 5. A linear mixed effects regression model was constructed to model log reaction time in correctly-answered target trials (i.e. trials where the answer is “yes”). The model structure and selection routine were the same as that of the accuracy model. Significance was determined by Markov chain Monte Carlo sampling (Baayen, 2008).

Prior to analysis, all target items with lower than chance accuracy (i.e. <50%) were removed, resulting in the removal of 6 target items (5.2% of total). A total of 1,074 data points were therefore included in the accuracy analysis. To remove extreme
outliers, all offset RTs more than 3 standard deviations away from each participant’s mean, and all RTs less than 100ms, were removed. 54 RTs were therefore removed, leaving a total of 943 RTs available for analysis. Due to equipment failure, one participant did not complete the AQ task and thus does not have an AQ score for analysis.

3.4.3 Results

3.4.3.1 Accuracy

The mean accuracy on target trials was 87.8%. The full model output for the final accuracy model is shown in Table 3.6. Simple effects of sentence context and talker AQ score were observed, although these effects are best considered in light of their interaction, which was also significant and is graphed in Figure 3.16. Overall, sentences with a biased context were more accurately responded to, and this effect was larger for sentences from high-AQ talkers than from low-AQ talkers.

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.33</td>
<td>-1.539</td>
<td>.124</td>
</tr>
<tr>
<td>Word duration</td>
<td>-2.202</td>
<td>-1.596</td>
<td>.110</td>
</tr>
<tr>
<td>Word production: LP</td>
<td>-0.021</td>
<td>-0.091</td>
<td>.928</td>
</tr>
<tr>
<td>Listener RMITE score</td>
<td>0.064</td>
<td>1.217</td>
<td>.223</td>
</tr>
<tr>
<td>Sentence context: Neutral</td>
<td>5.949</td>
<td>2.834</td>
<td>.005 **</td>
</tr>
<tr>
<td>Talker AQ score</td>
<td>0.067</td>
<td>3.251</td>
<td>.001 **</td>
</tr>
<tr>
<td>Listener AQ score</td>
<td>0.014</td>
<td>1.083</td>
<td>.279</td>
</tr>
<tr>
<td>Word duration × Word production:LP</td>
<td>-1.179</td>
<td>-0.758</td>
<td>.448</td>
</tr>
<tr>
<td>Word duration × Listener RMITE</td>
<td>0.587</td>
<td>2.375</td>
<td>.018 *</td>
</tr>
<tr>
<td>Word production: LP × Listener RMITE</td>
<td>0.041</td>
<td>0.739</td>
<td>.460</td>
</tr>
<tr>
<td>Sentence context: Neutral × Talker AQ</td>
<td>-0.075</td>
<td>-3.725</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>W duration × W production: LP × RMITE</td>
<td>-0.957</td>
<td>-2.742</td>
<td>.006 **</td>
</tr>
</tbody>
</table>

Table 3.6: Model output for sentence acceptability response accuracy.
A two-way interaction between word duration and listener RMITE score was also observed, along with a three-way interaction between word duration, listener RMITE score, and word production context. This interaction is difficult to interpret, but an attempt has been made in Figure 3.17. This figure was created by calculating mean accuracy rates per listener per word production context (HP or LP) per word duration (split into two equally sized bins). The difference between the HP and LP contexts was then calculated, resulting in 2 measures per listener of how much the HP–LP distinction helped or hindered their classification for words of differing durations. This measure is plotted on the y-axis of Figure 3.17 as a function of listener RMITE score, with the
results from each duration bin plotted separately. The panel labels depict the bin limits. By comparing both trend lines, a clear pattern emerges: the slope of the trend line becomes more positive, and the intercept lower, as the word durations increase.

In the sentences with words of shorter duration (left panel), HP words confer a small positive benefit to low-RMITE listeners, but not high-RMITE listeners (i.e. the slope is negative). In the sentences with words of longer duration (right panel), HP words confer a small cost to low-RMITE listeners, but, again, not high RMITE listeners (i.e. the slope is positive). This interaction suggest that for low-RMITE listeners, sentences with HP words were judged more accurately than those with LP words (HP>LP) when the word duration was short; but when the word duration was long, sentences with LP words were instead judged more accurately than those with HP words (HP<LP). The higher-RMITE listeners, on the other hand, were relatively unaffected by interactions between duration and word production context.

![Figure 3.17](image_url): Boost in sentence acceptability accuracy between HP and LP words as a function of talker RMITE score, binned by word duration in milliseconds. See text for details.
Finally, also present in the model but not involved in any significant effects or interactions was listener AQ score. This factor contributed significantly to data likelihood of the model ($\chi^2(1) = 24.326, p < 0.001$) despite its apparent lack of significance within the model itself.

### 3.4.3.2 Reaction time

The final, simplest model for reaction time is summarized in Table 3.7. Significant effects of final word duration, sentence context, and their interaction were observed. These effects are visualized in Figure 3.18, where it can be seen that, overall, neutral sentences were responded to much more slowly than biased sentences. Additionally, the graph shows that word duration confers an RT benefit for biased sentences, but an RT slowdown for neutral sentences. In other words, there is a processing benefit for the words of a longer duration, relative to the shorter words (cf. a similar effect in the lexical decision task above), but only when the sentence context is biased toward (or predictive of) the final word.

Although the effect of listener AQ was, again, not significant within the model, retention of the factor in the model significantly improved data likelihood ($\chi^2(1) =

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p_{MCMC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.411</td>
<td>-5.782</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Word duration</td>
<td>-0.711</td>
<td>-2.620</td>
<td>.011 *</td>
</tr>
<tr>
<td>Sentence context: Neutral</td>
<td>0.614</td>
<td>16.842</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Listener AQ</td>
<td>0.001</td>
<td>0.268</td>
<td>.737</td>
</tr>
<tr>
<td>W duration $\times$ S Context: Neutral</td>
<td>0.969</td>
<td>4.122</td>
<td>&lt; .001 ***</td>
</tr>
</tbody>
</table>

Table 3.7: Model output for sentence acceptability reaction times.
Figure 3.18: Reaction times in the sentence acceptability task as a function of final word duration, split by sentence context. Points indicate item mean and standard error.

48.672, \( p < 0.001 \). Figure 3.19 shows the influence of listener AQ on RT; overall there is a very small trend of slower responses by high-AQ listeners than low-AQ listeners, but the slope was not significantly different from zero.

### 3.4.4 Discussion

Overall, the results of this experiment provide limited support for the notion that unreduced speech facilitates semantic acceptability judgements, but with considerable caveats. The consideration of the linking hypotheses required for inference is an important step in interpreting the results. This task involved the manipulation of phonetic
properties of the final word in the sentence. In the processing of this final word, three steps must occur before a decision can be made: first, the word itself must be recognized and accessed; second, the word must be integrated into the rest of the sentence parse; third, the parse must be evaluated for sensicality. We know from the results of experiments 4 and 5 that the first step is harder and slower for short, reduced words relative to long, unreduced words. The second step ought to be equally easy for all target stimuli in this sentence, since all stimuli were nouns and every sentence context supported (indeed, required) a noun in the sentence-final position. A priori, we expect that the third step is easier for words in a biased context than in a neutral context,
since in the biased context, expectations have been set up about the likely semantic content or class of the final word. Given these assumptions, inaccurate responses (i.e. responses of nonsensicality) must then arise either at step 1, when the listener misperceives a word, or at step 3, when the listener cannot quickly resolve the meaning of the final word with the rest of the sentence. Although every target sentence was grammatical, many of them require the supposition of a reasonable discourse context, especially the sentences with neutral context. This supposition may be difficult to perform rapidly, leading to participants' rejection of stimuli.

In terms of accuracy, as predicted, sentences with biased contexts were far more accurately judged ($M = 96.3\%$) than sentences with neutral contexts ($M = 78.4\%$). Additionally, an interaction with talker AQ was observed such that the difference between neutral and biased contexts was larger for sentences spoken by high-AQ talkers, and smaller for sentences spoken by low-AQ talkers. Again bearing in mind that talker AQ score is a proxy for some unmeasured acoustic parameter(s) which varies systematically between talkers, this finding suggests that there is some acoustic difference in the production of the sentence contexts between talkers. That is, since the final word was balanced across the stimuli, this effect must be due to some acoustic distinction within the actual sentence material itself which helps or hinders the listener. Some aspect of the production of the neutral sentences by the high-AQ talkers leads to more inaccurate responses, while some aspect of their production of the biased sentences leads to more accurate responses. While this intriguing result is interesting in its own right, it does not bear directly on the present question of the role of reduced versus unreduced speech in sentence processing, and will not be discussed further.

The other observed effect on accuracy was a complex three-way interaction among the variables of word duration, word production context (LP or HP), and listener RMITE score. The result can be summarized as follows: for listeners with low
RMITE scores, sentences with short HP words were perceived more accurately than those with short LP words (HP>LP), but sentences with long HP words were perceived less accurately than those with long LP words (HP<LP). Again, the immediate interpretation of this effect is not clear, but the general trend is consistent with the results of experiments 4 and 5. Consider first that participants with low RMITE scores have poor theory of mind; consider also the fact that the HP–LP distinction in these models is a reflection of an unmeasured acoustic dimension not captured by duration. This acoustic dimension is not necessarily the same dimension in the words of short duration as in the words of long duration, and therefore we do not necessarily expect its effects on perception to be linear or straightforward. The listeners with the poor theory of mind are sensitive to this acoustic dimension(s), since it causes the differentiation in their accuracy rates, but the listeners with the better theory of mind (higher RMITE scores) are not. This listener-specificity is very similar to effects observed in experiment 5, whereby listeners with poorer ToM ability were more sensitive to acoustic cues or disturbances than others. It can be concluded, then, that there is considerable variation between individuals in how easily and effectively they respond to sentences with varying levels of acoustic clarity, and that this variability is partially linked to theory of mind ability. With that being said, it should be borne in mind that the HP–LP and neutral–biased distinctions were manipulated between-subject: while every listener heard all combinations of these factors, the comparison between the LP and HP productions of cake is necessarily between subjects. Although the four stimulus lists were balanced for number of participants (10 listeners in three lists and 9 in another), because individual variation in AQ and RMITE is not known at the time of a participant being assigned to a list, it is possible that the lists are unbalanced in the types of individuals that composed them. To test for this possibility, two one-way between-subject ANOVAs were conducted to predict AQ and RMITE scores from the experiment list (1
through 4, coded as categorical variables). List did not have an effect on either AQ score \( (F(3,34) = 0.533, p = 0.663) \) or RMITE score \( (F(3,35) = 1.873, p = 0.152) \), suggesting that these variables were not significantly unbalanced in their distributions across the lists. Nevertheless, there is still the potential that some of the observed effects arose due to coincidental combinations of listener tendencies with idiosyncrasies of the target tokens in a particular list.

In terms of reaction time, the largest effect was that biased sentences were responded to far faster \( (M = 537\text{ms}) \) than neutral sentences \( (M = 895\text{ms}) \). This result is not unexpected, since the neutral sentences are somewhat unlike ‘everyday’ sentences. The biased sentences also constrain the possible endings, thus allowing faster processing. For example, with For your birthday I baked a... the set of plausible endings is relatively small. After hearing the last word, the listener can check if the perceived word was in the set of plausible endings and make a decision accordingly. In contrast, the sentence Tom wants to know about the... has a much larger set of plausible endings, and the search takes a correspondingly longer amount of time.

Additionally, an interaction with duration was observed, such that biased sentences with longer words were responded to faster than those with shorter words. This effect is consistent with the results of experiments 4 and 5 which suggest that longer, unreduced words are processed faster and more accurately than shorter, reduced words. The other side of this interaction was that neutral sentences with longer words were responded to slower than those with shorter words. This unexpected effect is not consistent with prior results—it is not clear why unreduced words should take longer to process than reduced words. A possible explanation could be that the neutral context of the sentence acted as a kind of signal noise, or, put another way, the lack of a supportive context acts as noise relative to the helpful context of the biased sentences. Nevertheless, the influence of duration in neutral sentences is an unexpected result.
These observed results for both accuracy and reaction time cannot be attributed to oddities in the stimuli arising due to the splicing technique used to create them. All sentences where the sentence context matched the word production (i.e. neutral + LP; biased + HP) were unspliced, while sentences where there was a mismatch (i.e. neutral + HP; biased + LP) were spliced. If the listeners were able to perceive and respond to this splicing, or indeed if the listeners were sensitive to what could be considered a ‘natural’ sequence of sentence plus (un)reduced final word, then the modeling would have revealed a significant interaction between word production and sentence context. This interaction was absent from the final models of accuracy and reaction time, suggesting that the results cannot be attributed to that factor.⁸

### 3.5 General discussion

Taken together, the results of these experiments support the hypothesis that unreduced speech facilitates perception at multiple levels of processing. Experiment 4, which measured intelligibility and subjective clarity, established that longer, unreduced words were more intelligible and rated as more clear than shorter, unreduced words. No effects of listener variation in AQ or RMITE scores were observed. Experiment 5, a lexical decision task, showed that longer and more intelligible words were responded to more quickly and more accurately than shorter and less intelligible words. Additionally, two interactions with word production context suggested that, under certain conditions, words excised from a HP context, which are expected to be reduced, are responded to more slowly and less accurately than words excised from a LP context, which are

---

⁸Alternative methods of testing for this effect include splitting up the data into ‘spliced’ and ‘unspliced’ groups and testing for differences in the dependent variables. Two paired t-tests on subject means failed to reach significance for either accuracy \((t(38) = -0.441, p = .662)\) or reaction time \((t(38) = 0.2995, p = .766)\).
Perception task | Effect of temporal reduction
---|---
Clarity judgement | less clear ratings
Recognition in noise | lower accuracy
Lexical decision | lower accuracy, slower responses
Acceptability judgement | lower accuracy, slower responses

Table 3.8: Summary of effects of temporal reduction on the dependent variables in each of the perceptual tasks used in this chapter.

expected to be unreduced. Finally, experiment 6 provided limited support for this hypothesis in the form of a sentence acceptability task. It was found that among sentences which biased the listener toward a particular interpretation or conclusion (as in (6)), those with unreduced final words were responded to more quickly and more accurately than those with reduced final words. This effect was not observed in sentences with a neutral context (as in (5)). Moreover, experiment 6 also provided evidence that sensitivity to small acoustic detail, and its influence on perception, varies between listeners, and that this variation can be linked to theory of mind ability. There may be another dimension of cognitive-personality variance which better captures this effect, although such an investigation is left for future research. Table 3.8 provides a summary of the effects of temporal reduction (measured via word duration) on each of the dependent variables in the four tasks across the three experiments. Taken together, then, these experiments provide clear evidence in favor of the hypothesis that unreduced speech is perceptually easier than reduced speech, at the level of basic intelligibility and conscious judgement (experiment 4), in terms of lexical access and retrieval (experiment 5), and, to some degree, these effects percolate up to influence judgements of semantic goodness as well (experiment 6).
These findings serve to validate the assumptions of theories invoking ‘audience design’ as a mechanism for variability in phonetic reduction. Since unreduced speech facilitates retrieval at multiple levels of processing, the talker can make sure to produce unreduced speech in situations where it would help the listener, such as an unpredictable context, and the talker is free to produce reduced speech, thus conserving energy, in situations where the listener’s comprehension is facilitated by the context. In this sense, the present findings confirm results of previous studies which have suggested that reduced speech has negative consequences for perception, relative to unreduced speech (e.g. Bard and Anderson, 1994; Bard et al., 2000; Brouwer et al., 2012; Ernestus et al., 2002; Fowler and Housum, 1987; Ranbom and Connine, 2007; Tucker, 2011). The present findings also extend these previous results in two directions. The first direction of expansion is that of variation in stimuli. Most studies of reduction have looked at unsubtle, categorical reductions, such as /t/-flapping (Tucker and Warner, 2007) or segment deletion (Brouwer et al., 2013). The stimuli in the present study varied in much more fine-grained details such as vowel duration; a consistent result was obtained, suggesting that reduction effects are similar for both ‘large-scale’ and ‘small-scale’ acoustic effects. The second direction is that of levels of processing. Again, the majority of studies of reduction perception have focused on relatively low levels of processing, such as intelligibility (Ernestus et al., 2002; Picheny et al., 1985, 1986) or lexical decision (Ranbom and Connine, 2007). The present study has confirmed these previous results while also extending the findings to semantic processing, although the effect sizes in experiment 6 were smaller than those of experiments 4 and 5.

A somewhat unexpected finding in the present results is the presence of considerable variability in listener behavior, systematically related to listeners’ scores in the AQ and RMITE tasks, measures of autistic traits and emotional cognition, respectively. The majority of these results can be reconciled through appeal to the literature on the
broader autism phenotype: for instance, listeners with more autistic traits tended to attend more to acoustic cues at the expense of social, contextual, or lexical factors, and this difference in attention had consequences for reaction time and accuracy in both the lexical decision and the semantic acceptability task. This observed tendency is very much in line with conceptually similar findings about people with autism (e.g. Alcántara et al., 2004; Clopper et al., 2012, 2013; Frith and Happé, 1994) and neurotypical people with more autistic traits (e.g. Stewart and Ota, 2008).

However, the reconciliation of these results with the developmental psychology literature is not the same thing as their reconciliation with listener-oriented theories of phonetic reduction. Indeed, such theories do not predict the presence of the observed individual variation, and are not easily able to account for this variability. This failure is due to the listener-oriented account’s reliance on relative homogeneity across a speech community, which follows from its assumption of a functionalist communicative imperative. If, as the theory contends, talkers produce speech in such a way as to balance articulatory effort with listener comprehension, then individual variation in perception, such as that observed in the present study, should lead to fine-tuned interlocutor effects. That is, talkers should fine-tune their productions for the individual differences of every interlocutor. Although listener effects do exist, they are generally of a global and gross scale, and often do not actually help the listener (see Smiljanić and Bradlow, 2009, for review). Although some listener-oriented theories allow for a generic idealized interlocutor rather than the specific listener (Turk, 2010), such theories are still unable to reconcile the production data reviewed in Chapter 2. Alternatives to the listener-oriented accounts, which often situate talker behavior as purely a function of activation levels or lexical accessibility (Baese-Berk and Goldrick, 2009; Bard et al., 2000; Gahl et al., 2012), are similarly unable to provide accounts of individual variation in perception, although they could well require less modification than the
listener-oriented theories to do so. Thus, theories of phonetic reduction, both in terms of production and perception, must be modified to adequately account for differences in cognition between individuals.
Chapter 4

On the conceptual unity of reduction:

Reconciling theory and data

4.1 Introduction

This chapter addresses the conceptual unity of phonetic reduction, and seeks to answer the question of whether ‘reduction’ is a useful label that can be meaningfully applied to the outcomes of speech-influencing cognitive processes. Data from three corpora are examined to assess the hypothesis that different types of predictability lead to different patterns of reduction.

The primary focus in this chapter, as with the other chapters of this thesis, is predictability-based phonetic reduction. This term refers to a relationship between the predictability of some linguistic element (such as a word or syllable) and the phonetic production of that element. The generalization that holds for most predictability effects is that elements which are more predictable tend to be produced with less phonetic substance or prominence. This kind of relationship has been reported to exist for a wide variety of elements in a number of acoustic dimensions for a variety of definitions of ‘predictability’; see Clopper and Turnbull (2015) for review.
Several accounts of this phenomenon exist. As reviewed in Chapter 1, the accounts can be roughly categorized into ‘listener-oriented’, ‘talker-oriented’, and ‘passive’ perspectives on the ultimate origin of these effects. The ‘listener-oriented’ accounts (e.g. Aylett and Turk, 2004; Lindblom, 1990; Wright, 2004) claim that talkers fail to reduce in cases where reduction would hinder listener comprehension, and reduce in other cases to conserve effort. Patterns of reduction are therefore driven by explicit modeling of the interlocutor. The ‘talker-oriented’ accounts (e.g. Baese-Berk and Goldrick, 2009; Bard et al., 2000; Bell et al., 2009) claim that highly activated words are produced in a reduced way, and that activation follows from lexical frequency, contextual predictability, salience in the discourse, and similar properties. Patterns of reduction are therefore driven by cognitive architecture. Finally, the ‘passive’ accounts (e.g. Pierrehumbert, 2001a, 2002; Silverman, 2012) claim that there are no active forces driving reduction, but that reduction persists due to constraints on effective communication and language acquisition. Patterns of reduction are therefore epiphenomena arising from exemplar dynamics.

4.1.1 Reductionism in reduction

All three of the major theories can be said to be reductionist; that is, phonetic reduction is regarded as a unitary process which has a single explanation. Implicit in all of these theories is the notion that ‘phonetic reduction’ is a meaningful label which can be applied to the outcomes of the same process. Indeed, this assumption is widespread throughout the literature, and transcends theoretical boundaries (Ernestus, 2014). Within theoretical phonology, for example, lenition is usually described as a form of reduction or ‘weakening’ (Honeybone, 2008), and a similarly reductionist
pattern of thought is observed: a unified theory of lenition is regarded as a desirable goal for phonology (Bauer, 1988, 2008; Kirchner, 2001).¹

Nevertheless, recent work has brought such reductionist accounts into question. For example, reductionist assumptions about the speech production process have naturally related phonetic reduction and phonetic coarticulation as two sides of the same coin (e.g. Lindblom et al., 2009; Moon and Lindblom, 1994). However, careful analysis by Scarborough (2013) has shown that these two processes are to some extent orthogonal. Her experimental results suggest that vowels in nasal contexts in words in dense neighborhoods undergo both hyperarticulation (greater formant dispersion from center of vowel space) and greater nasal coarticulation (lower A1-P0 values) than vowels in words in sparse neighborhoods.² These results are not predicted under an account whereby reduction and coarticulation are underlyingly the same phenomenon. Instead, Scarborough (2013) argued that both vowel hyperarticulation and nasal coarticulation provide helpful cues to the listener regarding the identity of the word, cues necessitated by the dense neighborhood the words inhabit. She noted that speech communication relies on word identification, which consists of recognizing segments in context, rather than identifying individual segments in turn. Coarticulation, far from being a source of noise in the signal, instead aids in the recognition of segments in context (Beddor

¹It should be noted that there is nothing wrong with reductionism per se. The majority of the scientific enterprise is founded on (physicalist) reductionism, and Occam's razor dictates reductionist explanations as prima facie preferable hypotheses. However, the limits of reductionism have been acknowledged, particularly for complex biological domains such as ecology and behavior (see, for example, Bock and Goode, 1998). Ulanowicz (1996) outlined a number of aspects of the study of ecosystems which are not easily modeled through a reductionist or 'Newtonian' approach; many of these aspects, such as positive feedback loops and indirect mutualism, are true of the social and cognitive organization of language too. An emergent perspective, where complex phenomena are regarded as resulting from several interacting forces, is often preferable to explaining such systems (but see Bauchau, 2006, for a reconciliation of this common but false dichotomy).

²A1-P0 is the difference in amplitude between the first formant and the first low-frequency nasal peak in the spectrum, a common measure of vowel nasality (Chen, 1997).
et al., 2013). Regardless of the explanation, however, simple and uncontroversial assumptions about what features constitute ‘reduction’ ought to be questioned and tested.

More evidence for disunity in processes of reduction comes from a series of studies relating prosodic and temporal reductions to contextual predictability (Burdin et al., 2015; Turnbull, 2015; Turnbull et al., 2015). This work is reviewed in detail by Clopper and Turnbull (2015); the overall finding is that different acoustic dimensions are affected by different sources of predictability. For example, Turnbull (2015) observed second mention reduction influencing duration, but not F0, while some definitions of contextual predictability influenced both F0 and duration. Similarly, Burdin et al.’s (2015) comparison of focus-marking strategies in four typologically and genetically unrelated languages—Moroccan Arabic, American English, Paraguayan Guarani, and K’iche’—found language-specific patterns of reduction in prosodic prominence.

In addition to this disunity of reduction in production, there is also evidence that different kinds of reduction are processed differently in perception. Mitterer (2011) presented data showing that Dutch /t/-deletion and nasal place assimilation, two common processes of reduction, are processed by different mechanisms. This result suggests that there is not a single general mechanism for perceptual accommodation to these reductions. Hints of these distinctions were also present in early production work in laboratory phonology, such as Hewlett and Shockey’s (1992) finding that some coarticulatory patterns are affected by speech style more than others. Taken together with Scarborough’s (2013) and Turnbull’s (2015) research, it is clear that reductionist accounts of reduction may not be able to successfully account for the observed phenomena.
4.1.2 A hybrid model: the Multiple Source view

A reasonable response to this variability in reduction is the suggestion that perhaps a ‘hybrid’ model is appropriate: some phonetic reduction is listener-oriented, and some is talker-oriented. Watson’s (2010) ‘Multiple Source’ view is one such attempt to unify the disparate theories of phonetic reduction.

The ‘Multiple Source’ view was put forth by Watson (2010) as an account of the causes of prominence in speech. The literature and phenomena discussed by Watson (2010) make it clear that this theory is designed to account for phonetic reduction and enhancement as reflected in F0, intensity, and duration. The theory posits that different kinds of prominence relationships have different cognitive sources. Principally, Watson (2010) marshaled evidence to suggest that (in English) durational lengthening is generally tied to production difficulties, i.e. a talker-oriented effect, conceptually concordant with the proposals of Bell et al. (2009) and Fox Tree and Clark (1997). On the other hand, effects that influence F0 and intensity are claimed to signal changes to the discourse structure, and are thus listener-oriented, somewhat consistent with Turk’s (2010) arguments.

Unfortunately, Watson’s (2010) paper lacks a fully formalized theory and uses unclear terminology, which makes it difficult to evaluate. One of the claims made, for instance, is that prominence is a continuous, rather than categorical property. Watson paints the linguistic literature (e.g. Ladd, 2008; Schwarzschild, 1999; Selkirk, 1995) as regarding prominence as a categorical feature, which is true if ‘prominence’ is defined as a metrical or phonological feature—crucially, an abstract one. In this case, having evidence that prominence is continuous would indeed be an exciting and controversial result. However, both the evidence brought to bear and the terminology used suggest that we ought to infer that Watson’s (2010) conception of prominence is as an
acoustic phonetic property, and this view is stated explicitly on p170: “prominence is best understood as a continuous property of the speech signal”. From this perspective, prominence is trivially continuous since it is an acoustic feature of admittedly unknown dimensionality. This lack of clarity about what exactly prominence is makes it difficult to evaluate this model. Nevertheless, despite these difficulties, it is likely that a hybrid model like the Multiple Source view is necessary to account for phonetic reduction in speech.

4.1.3 Research question

As we have seen, the three major theories of predictability-based phonetic reduction can be said to regard reduction as a single phenomenon. Under this view, some hidden factor, which we will term PREDICTABILITY for convenience, is underlying the measures of frequency, neighborhood density, discourse mention, and semantic predictability. That is, each of these factors are not regarded as an end in themselves, but each is a manifestation of the underlying PREDICTABILITY present in the cognitive system. This reasoning sanctions the treatment of mathematically distinct calculations as measures of the same factor—for example, unigram, bigram, and trigram probabilities, conditional probabilities (calculated both forward and backward), and other factors (e.g. Seyfarth, 2014).

Alternatively, a hybrid model, such as Watson’s (2010) Multiple Source view, predicts that the sources of reduction are multiple. The four types of reduction considered in detail in this dissertation are frequency, neighborhood density, discourse mention, and semantic predictability. In the absence of clear predictions about each of

---

3 This quotation is one example of an apparent conflation of phonetics with phonology; earlier in the paper, Watson (2010) referred to “acoustic features underlying prominence” (p165) and “acoustic correlates of prominence” (p167). However, by definition, a property of the signal does not have acoustic correlates; it is acoustic.
these, a speculative hybrid model could propose that the lexical factors—frequency and neighborhood density—arise from one source, while the contextual factors—mention and semantic predictability—arise from another source. The goal of this chapter is to explore this alternative to a reductionist account through analysis of phonetic corpora, and, more generally, the conceptual unity of phonetic reduction.

4.2 Methodology: Generalized additive mixed models

The general modeling approach in this chapter involves the use of generalized additive mixed models (GAMMs) to explore the contributions of various predictability factors to speech production. These models are an extension of linear models, where the dependent variable is predicted through some smooth functions of the predictor variables, rather than the simple linear scaling of the predictor variables employed in linear models. This formulation allows for highly flexible models which are able to reflect with some accuracy the structure of the data, while still reflecting some underlying assumptions about how the data were originally generated. These models allow flexible investigation of covariates which may be strongly nonlinear in their effects on the dependent variable, and because the smoothing functions can take multiple arguments, the modeling of complex interactions becomes relatively straightforward. Two covariates which are believed to reflect a single process can thus be combined into a single term. At the same time, the models are essentially linear models at their core, and therefore the same tools for modeling and inference as are used in linear models can be used with GAMMs. Bearing in mind the complex and somewhat non-linear patterns of reduction observed in Chapter 2, GAMMs then provide an ideal means of modeling the effects of different lexical and contextual factors on phonetic reduction.
However, this flexibility comes at a cost, both practical and conceptual. Practically, GAMMs are computationally complex and expensive, with the calculation time rising exponentially with model complexity. Conceptually, inference from GAMMs is hampered by a number of problems such as reliable estimation of confidence intervals. Although the familiar tools of linear models can be applied, their interpretation must be tempered.

GAMMs, and additive modeling in general, have seen a steady rise in popularity and are particularly popular in the fields of ecology and plant science (e.g. Schmidt et al., 2011). In linguistics, GAMMs have been used primarily by Harald Baayen and colleagues (e.g. Baayen et al., 2011; Kösling et al., 2013; Shaoul et al., 2014; Wieling et al., 2011), but see also Plummer et al. (2013). As the theoretical results from mathematical statistics clarify and develop the underlying processes, and as the technical results from computational statistics develop more reliable algorithms and techniques, GAMMs are likely to become more widespread.

This section includes an elementary and somewhat simplified introduction to the mathematical foundations of generalized additive models, and their usage in statistical modeling and inference. This section attempts to capture, in broad strokes, what additive models do, and the conceptual basis for their construction. Details of their computational or algorithmic implementation are largely left aside; the interested reader is directed to Faraway (2006, chapter 12); Hastie et al. (2009, chapter 9); and Wood (2006, chapters 3–6).

An additive model is a special type of linear regression model. To begin, consider the basic linear regression model, where $y$ is the dependent variable, $x_i$ is a covariate (fixed effect), $\beta_i$ the coefficient, and $\epsilon$ the normally-distributed error term.

$$y = \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon$$  \hfill (4.1)
Generalized linear regression follows simply by transforming the dependent variable, as shown in equation 4.2.

\[ g(y) = \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon \]  

(4.2)

Here, \( g \) is some link function. The simplest case, \( g(x) = x \), is equivalent to a simple linear model. Other examples include \( g(x) = \log(x) \), which is used for Poisson models, and \( g(x) = \log\left(\frac{x}{1-x}\right) \), which is used for binomial models. Generalized linear regression models also require assuming that \( y \) is from some exponential family distribution; our three examples require normal, Poisson, and binomial distributions respectively.

A \textit{generalized additive model} is a generalized linear regression model where the dependent variable is estimated as a smooth function(s) of the predictors, as in equation 4.3.

\[ g(y) = f_1(x_1) + \ldots + f_n(x_n) + \epsilon \]  

(4.3)

The smooth functions \( f \) can take an arbitrary number of arguments; similarly, there is no restriction against strictly parametric model components. Therefore, model structures like in equation 4.4 are perfectly fine generalized additive models.

\[ g(y) = \beta_1 x_1 + \beta_2 x_2 + f_3(x_3) + f_4(x_4, x_5) + \epsilon \]  

(4.4)

In both 4.3 and 4.4, the functions \( f \) are composed of the sum of a set of \textit{basis functions} \( b \), where \( b_i \) is the \( i \)th such function, as shown in equation 4.5.

\[ f(x) = \sum \beta_i b_i(x) \]  

(4.5)

For a concrete example, consider a case where \( b \) is composed of cubic polynomials. The set of functions is shown in equation 4.6.

\[ b_1(x) = x^0, b_2(x) = x^1, b_3(x) = x^2, b_4(x) = x^3 \]  

(4.6)
Applying this basis to 4.5 yields 4.7.

\[ f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 \] (4.7)

The coefficients for these functions are fit using similar methods to the fitting of linear models in general, as described below. First, however, we will consider common basis functions used in additive modeling.

One of the more popular bases is thin plate regression splines. Full discussion of the mathematical foundations of this method is beyond the scope of this chapter, but see Wood (2003) for details. In general terms, thin plate regression splines estimate a smooth function of some number of predictor variables, and do so in such a way that minimizes the function being arbitrarily ‘wiggly’. They also have the useful property that they can be estimated without the ‘knots’ (turning points of the curve) being specified in advance, thus reducing subjective or arbitrary influence on the smooth function.

An important feature of thin plate regression splines is that the functions are optimized in such a way that the arguments to a given function are isotropic—that is, invariant to rotation in covariate space. In other words, a smooth function \( f(x, y) \) constructed with thin plate regression splines treats the variance in \( x \) equally to the variance in \( y \). This feature is very useful for certain applications, such as the modeling of geographic coordinates (Schmidt et al., 2011; Wieling et al., 2011) or appropriately transformed vowel spaces (Plummer et al., 2013). However, for other applications, including the consideration of lexical features such as frequency and density, this assumption is inappropriate. An alternative basis, advocated by Wood (2006), is tensor product smoothing. Again, the mathematical details are beyond the scope of this chapter, but Wood (2006, section 4.1.8) provides an overview. Briefly, in addition to the desirable property of anisotropy, tensor product smooths are invariant to linear rescaling of covariates. All else being equal, however, Wood (2006) recommends the use
of thin plate regression splines when isotropy is a reasonable assumption, or when the smooth is only to a single covariate. This recommendation is based on the fact that thin plate regression splines have been shown to be mathematically optimal under certain conditions (Wood, 2003), while theoretical results regarding tensor product smoothing have not yet been established. In this chapter, thin plate regression splines were used for all smooths of single predictors; smooths of multiple predictors used tensor products (Kösling et al., 2013, used a similar approach for the analysis of F0 contours).

As mentioned above, the parameters on the smoothing functions are selected in a similar manner to coefficients in linear models. The sum of least squares is the usual diagnostic for linear models:

\[ \sum_{i}(y_i - f(x_i))^2 \]  

(4.8)

In evaluating a smooth function \( f \), the sum of least squares is useful, but the best-fitting line is simply a cubic interpolation through the points. Such an interpolation essentially overfits the data and produces a very ‘wiggly’ curve. In order to penalize ‘wiggliness’, the concept of sum of squares is translated into consideration of curves. Thus, we consider the integrated square of the second derivative, in addition to the sum of squares as show in equation 4.9.

\[ \sum_{i}(y_i - f(x_i))^2 + \lambda \int_{a}^{b} f''(x)^2 dx \]  

(4.9)

Here, \([a, b]\) is the domain of \( f \) and \( \lambda \) is an unknown smoothing parameter which determines how much wiggliness is penalized. As \( \lambda \rightarrow \infty \), wiggliness is maximally penalized and \( f \) tends toward a simple linear function (a straight line), while when \( \lambda = 0 \), \( f \) is completely un-penalized and maximally wiggly.

The selection of \( \lambda \) could be arbitrary, but a principled approach is possible. In an ideal case, \( \lambda \) should minimize the difference between the estimated function \( f \) and

---

4Generalized linear models are evaluated on likelihood; such calculation requires maximum likelihood estimation, the details of which are complex and not germane to the present discussion.
the true function $\phi$, minimizing $M$ of equation 4.10.

$$M = \frac{1}{n} \sum_{i} (f(x_i) - \phi(x_i))^2$$  \hspace{1cm} (4.10)

However, since the true $\phi$ is unknown, an approximation must be used instead. The most common method is a cross-validation technique; the *ordinary cross validation score* (OCV) is defined as shown in equation 4.11, where $f^{[-i]}$ is the function fitted to all data except $y_i$. Each data point is left out in turn and the function fitted, and then the squared difference between the missing point and its predicted value is calculated. Next, an average over all the squared deviations is taken. This score can be shown to approximate $M$, and thus minimizing this score leads to the optimal $\lambda$ value.\(^5\)

$$OCV = \frac{1}{n} \sum_{i} (f^{[-i]}(x_i) - y(x_i))^2$$  \hspace{1cm} (4.11)

Generalized additive *mixed* models (GAMMs) are defined in a very similar way to mixed regression models (Baayen et al., 2008). For generalized mixed effects regression, a random term $Zb$ is added to equation 4.2, where $Z$ is the matrix of random effects and $b$ the matrix of random coefficients. This addition yields equation 4.12.

$$g(y) = \beta_1 x_1 + \ldots + \beta_n x_n + Zb + \epsilon$$  \hspace{1cm} (4.12)

For a GAMM, the same process is applied to equation 4.3, yielding equation 4.13.

$$g(y) = f_1(x_1) + \ldots + f_n(x_n) + Zb + \epsilon$$  \hspace{1cm} (4.13)

As before, GAMMs can take ordinary parametric covariates, and the smoothing functions can take an arbitrary number of arguments.

---

\(^5\)In most practical applications, the ordinary cross validation score is not used; rather, the generalized cross validation score (Golub et al., 1979) is used, which is computationally simpler and more appropriate as a model evaluation metric. Detailed discussion of this score is beyond the scope of this chapter; the interested reader is directed to Wood (2006), particularly sections 4.5.2 and 4.5.3, and Wang (2012) section 16.4. In the present analyses, the generalized cross validation score is used.
One remaining piece of machinery is required for the present chapter, which is the notion of a variable-coefficient model (Hastie and Tibshirani, 1993). This model is a type of GA(M)M where the smooth functions may be multiplied by some covariate. For example, a variable-coefficient GAMM could look like equation 4.14.

\[ g(y) = f_1(x_1)x_2 + Zb + \epsilon \]  

(4.14)

This method of combining covariates of interest essentially conditions the output of the smooth on the second covariate \((x_2)\) in the example above, and can be thought of as conceptually similar to an interaction term in a regular linear model.\(^6\) There are therefore two approaches to including more than one covariate in a single term—the variable-coefficient approach \(f(x_1)x_2\) and the multiple-argument approach \(f(x_1, x_2)\).

The variable-coefficient method is much simpler mathematically and computationally, and leads to a more restricted model. The conceptual difference between the two approaches can be appreciated by considering the kinds of question they answer—\(f(x_1, x_2)\) tells us how \(x_1\) and \(x_2\) work together to influence the response variable \(y\), while \(f(x_1)x_2\) tells us how \(x_1\) conditioned on \(x_2\) influences \(y\). Both approaches are useful in answering different questions, and both are used in the present study.

---

\(^6\)The conceptual similarity to interaction terms in regular linear models can be appreciated by considering a variable-coefficient GAMM:

\[ y = f_1(x_1) + f_2(x_1)x_2 \]

A simple linear model involving fixed effects of \(x_1\), \(x_2\), and their interaction can be reorganized to show the similarity in structure:

\[ y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \]

\[ = \beta_1 x_1 + (\beta_2 + \beta_3 x_1) x_2 \]

If \(f_2\) is strictly linear—i.e. \(f_2(x) = \beta_2 + \beta_3 x\)—then the last term in these models is equivalent.
4.3 Experiment 7: The non-independence of lexical frequency and neighborhood density

This experiment tested the contributions of lexical frequency and neighborhood density to phonetic reduction in a corpus of read speech derived from the experiments reported in Chapter 2.

4.3.1 Methods

4.3.1.1 Corpus

This speech corpus was derived from the results of experiments 1a, 1b, and 3, reported in Chapter 2. Each token in the corpus is a spoken target word. The corpus has tokens from 64 talkers. Acoustic measurements of stressed vowel duration and dispersion are included. Other variables in the corpus include each word’s lexical properties of number of lexical neighbors and lexical frequency, taken from the Hoosier Mental Lexicon (Nusbaum et al., 1984). See Chapter 2 for more details on the data collection methods. This corpus contains 11,474 vowel tokens.

The corpus also contains some variation in contextual properties of the circumstances of the word’s production. Namely, these productions varied in speech style (clear or plain), discourse mention (first or second), and semantic predictability (high or low). However, these manipulations were not balanced throughout experiments 1a, 1b, and 3, resulting in an uneven distribution of tokens. For example, half of the tokens from experiment 1b were clear speech, while half were plain speech. Experiments 1a and 3 did not have this manipulation, and it is not clear whether these tokens ought to be tagged as ‘plain’ speech or left underspecified. Both alternatives lead to problems in modeling—the former means that the plain tokens outnumber the clear tokens by
approximately 5 to 1; the latter means that speech style can only meaningfully be con-
sidered for the subset of tokens from experiment 1b, which precludes the purpose of
modeling from a large corpus in the first place. For these reasons, then, these context-
tual properties were not considered in the analysis.

4.3.1.2 Materials

The words in this corpus were drawn from the stimulus sets of experiments 1a, 1b,
and 3. The stimulus set of experiment 3 consisted of the same target words as used by
Burdin et al. (2014a,b), which were carefully selected to be balanced for neighborhood
density and frequency. However, the stimuli for experiments 1a and 1b were not chosen
with those variables in mind. Therefore, it is worthwhile to consider the relationships
among these variables once all these data have been pooled into the present corpus.

The corpus includes 294 word types; 275 of these are monosyllabic, 18 disyl-
labic, and one word has three syllables. Log frequency and number of neighbors were

Figure 4.1: Number of neighbors plotted as a function of log frequency with linear
trend overlaid (left panel), and as a function of number of phonemes (right panel).

---

7The trisyllabic word is botany, which regularly undergoes elision to a disyllabic [bə?ni].
significantly positively correlated \((p = .009)\), although this relationship is extremely weak \((r^2 = .024)\). Additionally, a word’s number of neighbors was negatively correlated with the number of phonemes in the word \((r^2 = .422, p < .001)\). Log frequency and number of phonemes were not correlated \((r^2 = .006, p = .187)\). The relationships between frequency and density, and number of phonemes and density, are depicted in Figure 4.1.

### 4.3.1.3 Analysis

Two GAMMs were constructed, one to model vowel duration and one to model vowel dispersion. Each GAMM had the same fixed effect structure: a tensor product smooth of lexical frequency and neighborhood density together, plus the same smooth multiplied by the number of phonemes in the word. A tensor product smooth was chosen since frequency and density are anisotropic variables, making a thin plate regression spline inappropriate. The number of phonemes was included as a variable coefficient in order to allow the model to account for effects due to word length. Therefore, the fixed effect structure of the models was as shown in equation 4.15.

\[
f_1(\text{frequency, density}) + f_2(\text{frequency, density}) \cdot N\text{phonemes} \quad (4.15)
\]

Random intercepts of talker and word identity were also included. All variables were z-score standardized before being entered into the model. Each term was evaluated for inclusion in the model via likelihood ratio testing, with terms being excluded if their inclusion failed to contribute significantly \((\alpha = .05)\) to data likelihood.

Before reviewing the results, it is worthwhile to consider the possible outcomes of the model(s) and what that would mean for frequency and density as they pertain to phonetic reduction. GA(M)Ms output the smooth functions used to model the dependent variable. The most common way to visualize a function with a single argument
is a line graph, with $x$ on the x-axis and $f(x)$ on the y-axis, which allows the curve to be plotted. For functions with two arguments, like those of this GAMM model, the function is not a curve, but rather a surface. A common visualization for surfaces is a heat map or contour plot, where $x$ and $y$ are plotted on the x- and y-axes respectively, and $f(x, y)$ is graphed as a color, indicating the ‘heat’ or ‘depth’, or through contours, like on a map. Alternatively, truly three-dimensional plots are possible, with $f(x, y)$ on the z-axis. Since three-dimensional graphs are difficult to interpret when they cannot be interactively rotated, and since the medium of dissertation publishing does not yet allow for interactivity,\(^8\) I have opted for the heatmap/contour plot approach in visualizing these surfaces.

The shape of these surfaces reveal the nature of the interactions between the covariates. The surface being essentially linear suggests that the $x$ and $y$ variables are independent: $x$ and $y$ operate orthogonally and combine in an additive manner. A linear surface would be essentially flat; the tangent to the surface is the surface itself. A nonlinear surface, with observable ‘hills’ and ‘valleys’, on the other hand, indicates that there is some dependence between the two variables. In the current case, we expect that frequency and density should depend on each other to a considerable degree, according to the understanding that both of these variables are reflections of the same underlying lexical property.

\(^8\)Although today, pop-up books are mainly marketed towards children, this was not always the case. Prior to the 18th century, pop-up (or ‘movable’) books tended to deal with academic topics, such as the *Chronica Majora*, which dates from the 13th century and allowed interactive exploration of the author’s notion of philosophical truth, and the 16th century *Cosmographia Petri Apiani*, which allowed the reader to divine astrological predictions. Perhaps pop-up graphs in academic journals would help with visualization of multidimensional data?
4.3.2 Results

4.3.2.1 Vowel duration

All terms in the model for vowel duration contributed to data likelihood. Therefore, the final model involved the smoothed frequency and density function \( F(3) = 8.777, p < .001 \) and the interaction of that smooth with number of phonemes \( F(6.475) = 9.096, p < .001 \).

Figure 4.2 shows \( f_1 \), the smooth function for density and frequency (top panel), and \( f_2 \), the smooth function for density and frequency interacting with number of phonemes (bottom panel) for vowel duration. These plots use both contour lines and coloration to indicate the predicted (standardized) vowel duration value. Viewing the contours as depicting a topographic map of a physical area may help with interpretation. Before interpreting these graphs, the contribution of each of the smooth functions to the predicted value should be considered. The function of frequency and density, the top panel of Figure 4.2, demonstrates the effects of frequency and density, independent of number of phonemes in a word. The lower panel shows additional effects due to number of phonemes in a word. To calculate a predicted value for a given word of density \( x \), frequency \( y \), and \( z \) phonemes we take the value of the top panel at point \((x, y)\) and add it to the value of the bottom panel at \((x, y)\) multiplied by \( z \), i.e. \( f_1(x, y) + f_2(x, y) \cdot z \). Since the output of \( f_2 \) is multiplied by the number of phonemes, the larger the absolute value of this output is, the greater the influence of the number of phonemes. The output of this function can therefore be interpreted as the size of the effect of number of phonemes on vowel duration. If this effect is not related to word frequency or density, the output should be constant—i.e. the same value regardless of the input, \( f_2(x, y) \sim 1 \). If the size of the effect is related to frequency or density, then
\(f_2(x, y) \sim x\) and/or \(f_2(x, y) \sim y\); that is, the output should vary. For \(f_1\), this function directly outputs the ‘baseline’ vowel duration prior to adding the effects (if any) of number of phonemes, and it can be therefore interpreted in a straightforward way. If word frequency and density are not relevant for vowel duration, the output of \(f_1\) should not vary.

The output of \(f_1\) is depicted in the top panel of Figure 4.2. Here, effects of both frequency and density can be observed. An increase in either frequency or density causes a decrease in the output value, depicted by the warmer colors and the contour lines. The top right of the graph depicts the lowest output values (the warmest shades) and the bottom left depicts the highest output values (the coldest shades). Independently of number of phonemes in the word, high frequency words have shorter vowels than low frequency words, and words with many neighbors have shorter vowels than words with few neighbors.

The shape of this surface is also meaningful. Note that the contour lines are evenly spaced, suggesting that the slope of the ‘hill’ is constant and that these effects are linear. There is a slight tendency, looking from right to left, for the contour lines to trend more closely to the vertical. This small trend suggests that frequency has a sharper influence on vowel duration in high density words relative to low density words. Nevertheless, effects of density and frequency are observed throughout the space.

The lower panel of Figure 4.2 depicts \(f_2\), the effect of number of phonemes on vowel duration size. The output of the function is always negative, suggesting that increasing the number of phonemes causes a decrease in vowel duration. This result mirrors the findings reported in Chapter 2. Additionally, the surface depicted in the panel is not flat, suggesting that neighborhood size and word frequency interact with number of phonemes. Considering first the right hand portion of the graph, the surface
where standardized density is greater than zero, a clear effect of frequency can be seen. The function outputs lower values (warmer colors) as frequency increases, suggesting that the effect of number of phonemes increases with word frequency. The fact that the contour lines are close to horizontal, particularly in the lower portion of the right side of the graph, demonstrates that density interacts only slightly with number of phonemes.

Turning now to the left hand portion of the graph, there is a large ‘plateau’ between the predicted values of $-0.2$ and $-0.4$, which occupies the majority of the space below standardized density of zero. This shape suggests that the effect of number of phonemes is of a constant size for the majority of words with a standardized density of less than zero.

Summarizing the patterns observed, there are main effects of frequency and density on vowel duration, such that words with higher frequency and words in dense neighborhoods have shorter vowels than words with lower frequency and words in sparse neighborhoods. Additionally, words with more phonemes have shorter vowels than words with fewer phonemes. The size of the effect of number of phonemes interacts with both frequency and density in a nonlinear fashion. For high-density words, high-frequency words exhibit larger effects of number of phonemes (leading to more vowel shortening) than low frequency words (leading to less vowel shortening). Frequency does not affect the effect size of number of phonemes for low-density words. Taken together, then, while neighborhood density and frequency may appear to be somewhat independent in their effects on vowel duration, when number of phonemes in a word is considered, density and frequency can be observed to be related in complex ways.
Figure 4.2: Contour plots of predicted vowel duration from $f_1$, the smoothed function of frequency and density (top panel), and $f_2$, the smoothed function of frequency and density conditioned on number of phonemes (bottom panel).
4.3.2.2 Vowel dispersion

For the vowel dispersion model, the interaction term with number of phonemes did not contribute significantly to data likelihood. The smooth of frequency and density was justified by likelihood ratio testing ($\chi^2(6) = 1189.4, p < .001$), but it did not reach significance within the additive model itself ($F(4.632) = 1.119, p = .344$). Figure 4.3 shows a contour plot of the smoothed function of density and frequency for vowel dispersion. Again looking at the plot as a map, as if we are gazing upon a hill from above, there is a clear downward slope as frequency increases, suggesting that more frequent words have less disperse vowels than less frequent words. Toward the center

![Contour plot of predicted vowel dispersion from the smoothed function of frequency and density.](image)

**Figure 4.3:** Contour plots of predicted vowel dispersion from the smoothed function of frequency and density.
of the plot there is a clear plateau between 0.1 and 0.2, before a sharper rise toward the bottom right. Overall, while a clear trend of frequency is observed, the effects of density are more complicated. An increase in density does not necessarily increase or decrease dispersion, but rather it is dependent on the word frequency. For example, consider the point \((0, 0)\) on the graph, with a predicted standardized dispersion of between 0.1 and 0.2. Moving to the point \((1, 0)\), that is, increasing the density, leads to a decrease in predicted dispersion (between 0 and 0.1). However, an equivalent increase in density in a different part of the space—for example, moving from \((-1, 0)\) to \((-1, 1)\)—instead increases the predicted dispersion by approximately 0.1. It would not have been possible to capture these nonlinearities among predictors through a strictly linear model.

### 4.3.3 Discussion

The results of this experiment confirm existing knowledge that frequency and density have an effect on vowel duration. Generally, high-density words have shorter vowels than low-density words, consistent with the findings of Gahl et al. (2012), and high-frequency words have shorter and less disperse vowels than low-frequency words, consistent with Gahl’s (2008) findings. The non-linearities found in the duration data suggest that frequency effects on words in dense neighborhoods are larger than frequency effects on words in sparse neighborhoods, as evidenced in the bottom panel of Figure 4.2, where the slope is much more steep toward the right of the graph (high-density words) than by the left (low-density words). Additionally, this interactive effect was enhanced on words with many phonemes. In the dispersion data, while there was no interaction with the number of phonemes in a word, frequency and density were again related in a nonlinear fashion.
Moreover, the use of additive modeling has revealed that the effects of frequency and density on vowel duration are related in a non-trivially non-linear way. Although other researchers have attempted to link together frequency and density into a single measure, such as Luce and Pisoni’s (1998) frequency-weighted neighborhood probability or Scarborough’s (2004) R metric, these transformations are strictly linear and cannot capture the variance observed in the present results.

Part of the reason for the non-straightforward relationships observed could be the segmental content of the words themselves. The literature on how speech sounds are affected by nearby sounds has a long history (Hillenbrand et al., 2001; Öhman, 1966; Stevens and House, 1963), and it is well-established that vowel dispersion and duration are both sensitive to the the place and voicing of the postvocalic consonant. Segmental content has specifically been proposed as a confound in neighborhood density studies by Holliday and Turnbull (2015) and Gahl (2015). Gahl’s (2015) re-analysis of Wright’s (2004) dataset found that after controlling for the influence of surrounding consonants on vowel formants, neighborhood density did not contribute significantly to formant prediction. Similarly, Holliday and Turnbull (2015) observed neighborhood density effects in Korean which were consistent with the English literature (such as Munson and Solomon, 2004; Wright, 2004), but the effects disappeared once segmental content was controlled for.

This analytic difficulty caused by segmental content is intrinsic to the study of both neighborhood density and frequency. Since these values are properties of lexical items, one must vary the lexical items used in order to contrast the values, thus opening up any study to the confounding effects of segmental content. Contextual predictability, on the other hand, allows for the lexical (and thus segmental) content to be held constant, while the context is varied. As mentioned, although the current corpus contained some manipulation of contextual predictability—namely, semantic predictability
and discourse mention—these variables were not balanced throughout the corpus and thus were not appropriate to include in the modeling. Experiment 8, however, considers a corpus where these properties were balanced, thus allowing for clear examination of phonetic reduction in a case where segmental content is fully controlled.

4.4 Experiment 8: Patterns of phonetic reduction in the Clopper corpus

The results of experiment 7 were revealing, but consideration of the manipulation of contextual variability was hampered by the between-subjects design of the source studies and therefore a full analysis of the contextual factors was not possible. Experiment 8 sought to firstly replicate the results of experiment 7 using a novel corpus, and secondly to investigate the confluence (or lack thereof) of effects of contextual predictability on vowel production. This experiment used a corpus of speech collected by Cynthia Clopper and colleagues, where the crucial manipulations of lexical and contextual properties were within-subjects.

4.4.1 The Clopper corpus

This speech corpus is derived from an experiment conducted by Cynthia Clopper and colleagues (Burdin and Clopper, 2015; Burdin et al., 2014a,b). Participants read 30 short paragraph-length stories, constructed in the style of Baker and Bradlow’s (2009) stimuli. Each paragraph contained several target words which varied in the lexical properties of frequency and neighborhood density. Each target word also varied in the contextual properties of semantic predictability (measured in three ways, discussed below), and mention (first or second). Target words were generally monosyllabic nouns,
and are listed in full in Appendix E. Participants first read the 30 stories in a plain speech style—as if talking to a friend—and then read the stories again, in a different order, in a clear speech style—as if talking to someone with a hearing impairment or who is a non-native speaker of English. The recordings were automatically aligned with the Penn Forced Aligner (Yuan and Liberman, 2008) and vowel boundaries of the target words were hand-corrected. Vowel duration and first and second formant frequencies of the target words were measured, for a total of 28,185 vowels.

4.4.1.1 Semantic predictability measures

The corpus presents three measures of semantic predictability: cloze, bigram, and trigram probabilities. The cloze probabilities were derived from the results of a cloze task using the stimulus sentences. Sentences were orthographically presented to participants, with one (target) word missing, and participants were instructed to ‘fill in the blank’ with the most appropriate word. The proportion of times that participants answered correctly for a given word in a given context is the cloze probability. This task was completed by 101 participants from the same population as those who carried out the reading task. The bigram and trigram probabilities were conditional probabilities derived from the Google Web 1T corpus (Brants and Franz, 2006) using the Get 1T querying software (Hawker et al., 2007). Since the 1T corpus consists of frequency counts of word strings, transforming the frequencies into probabilities requires some computation, the details of which are provided here in full for reference.

The general formula for a conditional n-gram is provided in equation 4.16, where \( C(x) \) is the frequency count of the word or word string \( x \). In other words, the estimated probability of the word \( w_n \) given a preceding string \( s \) is the frequency of the string—i.e. \( C(s, w_n) \)—divided by the total frequency of all other strings of equal
length that begin with \( s \).

\[
\hat{p}(w_n|w_1, \ldots, w_{n-1}) = \frac{C(w_1, \ldots, w_n)}{\sum_j C(w_1, \ldots, w_j)} \tag{4.16}
\]

To reduce computational complexity (and thus decrease the querying time required to obtain the counts), the following simplifying assumption was made:

\[
\sum_j C(w_1, w_j) = C(w_1) \tag{4.17}
\]

That is, it was assumed that, for example, the frequency of all three-word phrases beginning with “the big” was equal to the frequency of the phrase “the big” alone. This assumption led to much faster corpus search times.

However, given the infinitude of language, there are invariably legitimate \( n \)-grams which are not attested in a corpus, even in large datasets like the Google Web 1T corpus. Rather than assign these unattested \( n \)-grams a count (and thus a probability) of zero, a smoothing modification was applied. The ‘smoothing’ of language engineering is not related to the ‘smooth functions’ of additive models discussed above. In smoothing, the probability mass of the unattested forms is increased, thus ‘smoothing’ the distribution of probability mass. Laplace Smoothing, also known as ‘add-one’ smoothing, was used, where all \( n \)-gram counts (both attested and unattested) were increased by 1.\(^9\) Therefore, the modified counts are easily defined:

\[
C^*(w) = C(w) + 1. \tag{4.18}
\]

To apply Laplace Smoothing, equation 4.18 can be applied to the numerator of equation 4.16. To smooth the denominator, however, equation 4.18 must be applied to every possible trigram sequence that fits the template, even the unattested ones. Given

\(^9\)Laplace Smoothing is inadequate for the majority of NLP engineering applications (Gale and Church, 1994); however the Google 1T corpus is too large to feasibly build a standard backoff (e.g. Katz) model, and Laplace Smoothing is easy to implement and understand.
V items in the vocabulary (defined as the total number of unique unigrams in the corpus\textsuperscript{10}), the smoothed denominator $D^*$ can be defined as in equation 4.19, where $D$ is the original, unsmoothed denominator.

$$D^* = V + D,$$ \hspace{1cm} (4.19)

Thus, substituting equations 4.18 and 4.19 into 4.16, the smoothed general equation is obtained:

$$\hat{p}^*(w_n|w_1, \ldots, w_{n-1}) = \frac{C^*(w_1 \ldots w_n)}{V + \sum_j C(w_1, \ldots, w_j)}$$ \hspace{1cm} (4.20)

Using this equation, bigram ($n = 2$) and trigram ($n = 3$) probabilities were calculated.

### 4.4.1.2 Measurement

Formant values at the vowel midpoint were converted to ERB and a measure of vowel dispersion was calculated. Dispersion was defined for each token as the Euclidean distance from that token to the talker-specific vowel center. The vowel center was defined as the mean of by-vowel means for that talker. Dispersion and duration values more than two standard deviations away from a talker’s mean values were excluded; this method resulted in the exclusion of 1,011 dispersion values and 830 duration values.

### 4.4.2 Replication of experiment 7

First, the analysis from experiment 7 was repeated, to assess the comparability of the two datasets.

\textsuperscript{10}This simplifying assumption was made for convenience, aided by the fact that the 1T corpus is so large. An alternative to this assumption, and appropriate for the modeling of smaller corpora, would be to use an independent source, such as a dictionary, for the number of unique wordforms in the language.
Figure 4.4: Contour plots of predicted vowel duration from smoothed functions of frequency and density (top panel) and frequency and density conditioned on number of phonemes (bottom panel).
4.4.2.1 Analysis

A GAML was constructed with a tensor product smooth of frequency and density, plus
that smooth multiplied by the number of phonemes in a word, and random intercepts
of word and talker. Variables were z-score standardized before being entered into the
analysis.

4.4.2.2 Results

Both terms in the vowel duration model contributed to data likelihood and were re-
tained. Figure 4.4 shows contour plots of the smooth functions of frequency and den-
sity, and frequency and density conditioned on number of phonemes. The topography
of the top panel shows warmer shades, depicting shorter vowel durations, tending to-
ward the right hand side of the graph, suggesting that words with more neighbors have
shorter vowels. Additionally, an effect of frequency can be observed toward the left
hand side of the graph: more frequent words tended to be produced with shorter vowel
durations than less frequent words, for words with a standardized density less than 1.
The bottom panel suggests that, like in experiment 7, words with more phonemes have
shorter vowels than words with fewer phonemes. Additionally, an interaction with fre-
quency was observed such that the effect of number of phonemes is diminished for
high-frequency words relative to low-frequency words.

The topography of these graphs are roughly similar to those of experiment 7's
Figure 4.2. Both reveal effects of frequency—more frequent words have shorter vowels
than less frequent words—and density—words in denser neighborhoods have shorter
vowels than words in sparse neighborhoods. There are two main differences, however.
The results of experiment 7 suggested that frequency effects were robustly observed
in words in dense neighborhoods, but not so much in words in sparse neighborhoods;
furthermore, this effect was increased in magnitude for words with many phonemes. In the current replication, almost the opposite pattern is observed: frequency effects were robustly observed in words in sparse neighborhoods, but not so much in words in dense neighborhoods; and this effect was decreased in magnitude for words with many phonemes. The interaction with number of phonemes is particularly striking: observe that the ‘bottom of the hill’ in the lower panel of Figure 4.2 is toward the bottom of the graph, while the ‘bottom of the hill’ in the lower panel of Figure 4.4 is toward the top of the graph.

This discrepancy between the data collected in this dissertation and Clopper’s data was noted in Chapter 2 during the discussion of the results of experiment 4. Experiment 4 and Clopper’s experiment used the same set of target words as stimuli, and experiment 4 comprises a portion of the corpus used in experiment 7. Therefore, the target words considered in the Clopper corpus constitute a proper subset of the target words considered in experiment 7, implying that the differences here cannot be attributed to accidental properties of the stimuli. The main difference was in elicitation method—Clopper’s experiment was speech read in paragraph contexts, while experiment 4 was simply a read word list. That this methodological difference would create such stark differences in the effects of frequency and density is unexpected.

In the vowel dispersion model none of the terms significantly contributed to data likelihood. Recall that the frequency and density smooth in the vowel dispersion model for experiment 7 did not reach significance within the model itself. It is likely that any effects on dispersion are relatively small, leading to the lack of significant results.
4.4.3 Evaluating the predictability variables

A priori, it is reasonable to expect that, of the three predictability variables, the cloze probabilities should reflect more closely the underlying cognitive representations of predictability. The 1T corpus is derived from written text on the web, and although the speech in the Clopper corpus consists of paragraphs read aloud, the genre of the web is somewhat different from spoken language. Moreover, experimental evidence suggests that cloze probabilities are better able to model behavioral data than n-gram probabilities (Smith and Levy, 2011). However, analyses of Clopper’s corpus have thus far failed to observe an effect of cloze probability on either vowel duration or vowel dispersion (Burdin and Clopper, 2015; Burdin et al., 2014a,b). Table 4.1 shows a correlation matrix of all three of these measures (log-transformed), confirming that they are significantly correlated.

<table>
<thead>
<tr>
<th></th>
<th>cloze</th>
<th>trigram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>(p)</td>
</tr>
<tr>
<td>trigram</td>
<td>.396</td>
<td>(&lt;.001)</td>
</tr>
<tr>
<td>bigram</td>
<td>.335</td>
<td>(&lt;.001)</td>
</tr>
</tbody>
</table>

Table 4.1: Correlation matrix of the three (log-transformed) word predictability measures in Clopper’s corpus.

Overall, words with higher density tended to have shorter vowels, and complex, non-linear interactions between density and frequency were observed. These interactive effects underscore the conclusion of experiment 7: that frequency and density are not independent parameters. Effects on dispersion, if present, were small.
4.4.3.1 Analysis

Given this choice of three predictability measures, it is not immediately obvious which one best reflects variation in phonetic reduction. To pit these measures against each other, several linear mixed effect regression models were constructed.\textsuperscript{11}

The design of the current analysis was motivated by several aspects of the design of the Clopper corpus. Recall from the earlier description of the corpus that half of the speech in the corpus was produced in a plain style, while the other half was produced in a clear style. Any analysis will have to either consider only a subset of the corpus, or include style as a factor in the modeling. Since each style represents the same talkers reading the same materials, there are no difficulties with nesting of variables. However, the situation of the contextual variables of mention and predictability are somewhat more complex. Each target word was produced twice in each paragraph, as a first and a second mention. Each word in context also has a unique predictability value. This melding together of predictability and mention has the consequence that comparisons within word type are complicated. It is possible to compare a given target word’s first mention tokens with its second mention tokens to assess effects of second mention reduction, but such a comparison is compromised by the fact that these words have different predictabilities. It is also possible to consider the effect of predictability by comparing the high-predictability token of a given word with its low-predictability token, but this comparison is again compromised by the fact that the tokens have different mention statuses. To get around these problems, the analysis considered relative differences in word pairs, rather than absolute measures of individual word tokens.

\textsuperscript{11}For this analysis, linear mixed effect modeling is being used rather than GAMMs. A preliminary attempt at a GAMM model for this analysis failed to converge.
First, the corpus was reorganized such that each data point was not an individual word production, but an individual word pair of the first and second mention. Such a reorganization has two primary advantages: one, it makes for a smaller and simpler data set (and therefore simpler modeling); two, it allows us to ignore lexical factors such as frequency and neighborhood density. Each data point is comparing a word with itself in a different context, and lexical factors are therefore kept constant. For each word pair, the difference between vowel duration and vowel dispersion in first and second mention was calculated. A positive value indicates that the first mention was longer or more disperse than the second mention, i.e. that second mention reduction had occurred. This restructuring led to 13,049 duration difference measures and 12,937 dispersion difference measures. Also calculated were the difference between the log transformed\(^{12}\) predictability measures between first and second mention. Here, a positive value indicates that the first mention was more predictable than the second mention. In other words, the predictability difference value is a measure of how predictable the first mention was relative to the second mention.

To assess, somewhat trivially, whether second mention reduction occurred in this corpus, one-sample \(t\)-tests (where \(\mu = 0\)) of subject means and word means were carried out on the duration and dispersion difference measures. The results indicate that the duration difference was significantly different from zero, by subject (\(t(29) = 4.286, p < .001\)) but not by word (\(t(230) = 1.400, p = .163\)), and that the dispersion difference was significantly different from zero by subject (\(t(29) = 8.756, p < .001\)) and by word (\(t(230) = 5.373, p < .001\)). Vowels in second mentioned words were an average of 2.1ms shorter and 0.065ERB less disperse than the first mention.

\(^{12}\)The predictability measures were log-transformed due to their skewed, non-normal distributions. Probabilities of zero (only attested for the cloze measure, due to the smoothing employed for the \(n\)-gram measures) were transformed to \(-25\) rather than \(-\infty\).
Two ‘baseline’ models were created, one predicting duration difference and the other dispersion difference. Speech style (clear or plain) was the only fixed effect. A random intercept of talker with style as a random slope, and a random intercept of word with style as a random slope were also included. Next, six ‘test’ models were created, each with the baseline fixed effect structure plus one of the predictability difference variables, and an interaction between that variable and style. Each test model was compared to the baseline by likelihood ratio testing. All variables were centered around the mean before being entered into the model.

### 4.4.3.2 Predictions

Since this modeling is somewhat conceptually complex, it is worthwhile considering the predicted outcomes. For ease of reference, difference in predictability (any kind) will be referred to as $\Delta p$, and difference in duration will be referred to as $\Delta d$. The predictions for dispersion are the same as that of duration, and $\Delta d$ can be read as difference in dispersion if desired. As described above, $\Delta d$ is a measure of second mention reduction—or, put another way, the duration of the word’s first mention ($w_1$) relative to the duration of its second mention ($w_2$). A positive $\Delta d$ indicates that $w_1$ is longer than $w_2$. $\Delta p$ is defined similarly and allows similar reasoning: a positive $\Delta p$ indicates that $w_1$ is more predictable than $w_2$. It is expected that more predictable words undergo reduction relative to less predictable words. Therefore, in word pairs with high $\Delta p$, it is expected that the duration of $w_1$ will be relatively shorter than the duration of $w_2$. That is, in word pairs with high $\Delta p$, we expect to observe a lower $\Delta d$, relative to pairs with a lower $\Delta p$. A negative correlation between $\Delta p$ and $\Delta d$ is predicted. Since second mention reduction was observed overall, the $\Delta d$ values in high-$\Delta p$ cases will not necessarily be less than zero, but simply lower than in low-$\Delta p$ cases.
Finally, an interaction with style is expected only if predictability and mention effects operate in different ways in clear speech than they do in plain speech. Such a result would speak to some degree of conscious control over these effects. A result of this type was reported by Baker and Bradlow (2009), who investigated interactions on word duration among frequency, mention, and speech style. They reported that high-frequency words had a larger degree of second mention reduction than low frequency words, but only in plain speech. That is, in clear speech, the difference between first and second mentions was not mediated by frequency. Baker and Bradlow (2009) took this result to mean that talkers reduce the most when all factors support reduction (frequency, mention, and style), and that there is a limit on word duration reduction in clear speech. Baker and Bradlow’s (2009) study was somewhat different from the current experiment, in that they examined word duration as a function of frequency, while the current experiment examined vowel duration and dispersion as a function of semantic predictability.

4.4.3.3 Results

For the duration difference models, the cloze predictability model was not significantly different from baseline ($\chi^2(2) = 5.671, p = .059$), but both the bigram ($\chi^2(2) = 2297.3, p < .001$) and the trigram ($\chi^2(2) = 15332, p < .001$) models were. Similarly, for the dispersion difference models, the cloze predictability model was again not significantly different from baseline ($\chi^2(2) = 1.813, p = .404$), but both the bigram ($\chi^2(2) = 340.67, p < .001$) and the trigram ($\chi^2(2) = 2650.8, p < .001$) models were.

For both dependent variables, the trigram models had higher log likelihood than the bigram models (vowel duration: $−57,382$ versus $−63,900$; vowel dispersion: $−9,701$ versus $−10,856$), and so were singled out for further investigation. The coefficients of the trigram models are reported in Table 4.2. The reported $p$-values were
calculated by treating the absolute $t$-statistic as if it came from a $t$-distribution with degrees of freedom equal to the number of data points minus the number of model parameters (Baayen, 2008).

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.545</td>
<td>-0.916</td>
<td>0.360</td>
</tr>
<tr>
<td>Trigram difference</td>
<td>0.221</td>
<td>0.469</td>
<td>0.639</td>
</tr>
<tr>
<td>Style: plain</td>
<td>-0.033</td>
<td>-0.040</td>
<td>0.968</td>
</tr>
<tr>
<td>Trigram $\times$ Style</td>
<td>-0.215</td>
<td>-0.909</td>
<td>0.363</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.011</td>
<td>-0.744</td>
<td>0.457</td>
</tr>
<tr>
<td>Trigram difference</td>
<td>-0.009</td>
<td>-2.433</td>
<td>0.015*</td>
</tr>
<tr>
<td>Style: Plain</td>
<td>0.008</td>
<td>0.750</td>
<td>0.453</td>
</tr>
<tr>
<td>Trigram $\times$ Style</td>
<td>0.000</td>
<td>-0.090</td>
<td>0.928</td>
</tr>
</tbody>
</table>

Table 4.2: Output for the trigram models for duration difference and dispersion difference in the Clopper corpus.

As can be seen in Table 4.2, none of the effects in the word duration model reached significance, despite the model’s improvement in data likelihood over the baseline model. This situation can be interpreted as the trigram difference score lending general predictive support to the model, without the coefficient being significantly different from zero. The dispersion model, however, showed a significant effect of the trigram difference, such that $\Delta p$ was negatively correlated with $\Delta d$, as predicted. This effect is depicted in Figure 4.5, where the effect can be clearly observed.

4.4.4 Discussion

This experiment has replicated the general results of experiment 7, namely, that the effects of frequency and density on vowel production are not independent and are
Figure 4.5: Dispersion difference in ERB between first and second word mentions as a function of the difference between log conditional trigram probability of the word pair. Points depict word means and standard errors.

involved in a complicated relationship. As discussed, it is not clear that this lack of independence can be addressed by linear transformations to density measures, such as Scarborough’s (2004) R or various conceptions of frequency-weighted number of neighbors (Luce and Pisoni, 1998).

The analysis of the contextual predictability variables suggested that, consistent with prior literature, more predictable words were produced with less disperse vowels than less predictable words. Somewhat surprisingly, no effect on vowel duration was noted. Additionally, the trigram probabilities were the best of the three predictability measures for modeling effects of predictability on reduction. Work on the Clopper corpus to date has used cloze probabilities and thus far failed to find strong effects
of predictability on vowel reduction or dispersion.\textsuperscript{13} This finding of trigram effects is therefore a novel one. Some aspect of the analytical setup may have facilitated this discovery. Firstly, the current analysis focused on word pairs, comparing like with like. Secondly, the predictability measures were rescaled to a relative measure. These two steps allowed for significant conceptual abstraction away from acoustic effects caused by segmental content and other extraneous factors.

Nevertheless, while it is not surprising that the trigram probabilities outperformed the bigram probabilities, it is quite odd that they also outperformed the cloze probabilities. Indeed, the cloze probabilities showed no relationship to vowel duration or dispersion. Given that the cloze probabilities reflect sentence-specific judgements about what word is appropriate in a particular position, and that the \textit{n}-gram probabilities were taken from a somewhat genre-inappropriate but massive database, this result is surprising. It is not common for computers to outperform humans at language tasks.

4.5 Experiment 9: Phonological reduction in the Buckeye corpus

4.5.1 Introduction

All of the reductions considered thus far in this thesis have reductions in what Hawkins (2003) termed “fine phonetic detail”; that is, small, gradient changes that do not substantially alter the realization of a word. For example, the second mention reduction observed in experiment 8 had an effect size of 2ms in vowel duration difference. These distinctions in fine phonetic detail may be consistent and robust, but listeners are not

\textsuperscript{13}However, see Burdin and Clopper (2015), who found that words with high cloze probabilities were less likely to be realized with a pitch accent than words with low cloze probabilities.
necessarily consciously aware of them (but cf. the results of Chapter 3, suggesting that these minute differences can and do influence various processes of speech perception).

However, the listener-oriented account of predictability-based phonetic reduction, in addition to the talker-oriented and passive evolutionary accounts, is concerned with reduction in a general sense. This final experiment therefore considers much larger-scale reductions, namely the wholesale deletion of segments. This kind of deletion can be considered to be phonological or categorical, as opposed to the ‘phonetic’ or gradient nature of the other reductions considered in this thesis. Segment deletion is necessarily consciously accessible to listeners, since the usual methods for its detection involve phonetic transcription. The purpose of this experiment was to determine if the same linguistic factors that influence reduction in fine phonetic detail also influence segment deletion.

Johnson (2004) identified a number of ‘massive reductions’ in a subset of the Buckeye Corpus of Spontaneous Speech where words had several phonemes deleted in production, such as \([\text{pær]}\) for *apparently*, \([\text{hle\_es]}\) for *hilarious*, and \([\text{ptik\_e}]\) for *particular*. Johnson (2004) observed greater reduction of (segmentally) longer words than shorter words, and higher rates of deletion in function words relative to content words. The current experiment extends Johnson’s (2004) work and considers the roles of lexical frequency, neighborhood density, and discourse mention in mediating segment deletion in the Buckeye Corpus.

Although it is reasonable to expect that segment deletion leads to a decrease in overall word duration, this relationship does not necessarily hold. It is possible for a word with a deleted segment to have the same or longer duration as a word with no deleted segments; and it is possible for a word with no deleted segments to undergo temporal reduction. These distinctions are not merely curiosities of production, but are potentially relevant for perception too. Ernestus and Baayen (2007) presented
evidence suggesting that, while segment deletion delays word recognition, shorter durations may speed up word comprehension. Therefore, while it may be tempting to expect that the conditions that give rise to segment deletion also necessarily lead to temporal reduction, the reality of the situation may be considerably more complex.

4.5.2 Methods

4.5.2.1 Corpus

The Buckeye Corpus of Conversational Speech (Pitt et al., 2007) is an orthographically and phonetically transcribed audio corpus of roughly 40 hours of speech. The corpus was collected by individually interviewing 40 talkers from central Ohio and asking them questions about their lives. Therefore, the corpus consists mostly of spontaneous monologues in a conversational style. The audio recordings were transcribed in terms of standard orthography, (dictionary) phonemic transcriptions, and broad phonetic transcriptions.

4.5.2.2 Analysis

Since each word in the corpus has both a phonetic and phonological transcription, it is possible to compare the two transcriptions and determine if any phonemes were deleted in the word’s pronunciation. For example, in the pronunciation of Columbus /kɔləmbəs/ as [klambəs], one phoneme (/ɔ/) was deleted. Some substitutions were tolerated: namely, the transformation of liquids to syllabic liquids, and alveolar stops to flaps. Therefore, the production of getting /gɛtɪŋ/ as [gɛrɪŋ] did not count as a deletion. Similarly, the pronunciation of holding /hoʊldɪŋ/ as [houn] involved the deletion of 3 phonemes: /l, d, t/. The substitution of [ŋ] for /ŋ/ was not regarded as a deletion.
For each word token, the number of deletions was calculated. The identity of the deleted phonemes was also recorded. The neighborhood density of each word was extracted from the Hoosier Mental Lexicon (Nusbaum et al., 1984), the frequency each the word was calculated from the rest of the Buckeye corpus, and the mention status (first, second, third, … nth) for each word in each conversation was calculated by counting how many times a word was uttered by the talker in each conversation. Finally, each word was tagged for whether it was a function or content word by using the CLAWS part of speech tagger (Garside, 1987; Garside and Smith, 1997).

Using GAMMs on this dataset was unfortunately computationally intractable due to its size. Therefore, a Poisson mixed effects regression model was constructed to predict the number of phoneme deletions per word. Fixed effects were the number of phonemes in the word, the log frequency of the word, the word's density, the log number of mentions, and the word’s class (function or content word). All possible two-way interactions between frequency, density, mention, and class, and all two- and three-way interactions between number of phonemes and the other effects were included, as shown in 4.21.

\[
N_{\text{phonemes}} \times (\text{frequency} + \text{density} + \text{mention} + \text{class})^2
\]  

(4.21)

From this maximal model, fixed effects were systematically removed and assessed for inclusion by likelihood ratio testing. Only effects which significantly contributed to data likelihood were retained. The inclusion of an interaction required the inclusion of the

14The corpus is large for a speech corpus, but relatively small for frequency estimations. At 282,435 words, it is a little over a quarter the size of the Brown corpus. However, estimates were taken from this corpus, rather than an external source, due to concerns about the specific genre of this corpus. Namely, since the majority of the conversations are about the talkers’ lives in central Ohio, some words (particularly place names such as Columbus and Dublin) are of much higher frequency than would be otherwise estimated by an external corpus.

15I am grateful to Jordana Heller for sharing these annotations with me.
component simple effects. In all models, random intercepts of speaker and word were included. Random slopes were not included to keep computation time reasonable.

4.5.3 Results

4.5.3.1 Descriptive trends

Before presenting the results of the modeling, we will first consider general trends in the data. Table 4.3 lists the percentage deletion rate for each phoneme. These data were derived by splitting up the corpus into a phoneme-level database. Here, each of the 904,498 phonemes represented in the corpus was tagged with whether or not it was deleted. Some general trends emerge in Table 4.3: voiced obstruents delete at much higher rates than their voiceless counterparts, with the exception of the pair /k, g/;\(^{16}\) consonants delete more often than vowels; of the ten most deleted phonemes (in order: /d, ð, t, v, ʃ, z, ɹ, dʒ, θ/), eight are coronals. These trends accord with reported trends of deletion in conversational speech (Johnson, 2004; Shockey, 2003).

4.5.3.2 Number of deletions per word

The optimal model from the model selection procedure is shown in Table 4.4. As expected, an effect of number of phonemes was observed, such that words with more phonemes are likely to have more deletions, concordant with Johnson’s (2004) findings. This effect follows from simple logic—there is a hard limit on the number of possible deletions for a given word, and that limit is larger for longer words than for

\(^{16}\)This disparity appears to be driven largely by two forces. First, /k/ is a common target for deletion in the high-frequency words think and can, with /k/-deletion rates of 18.5% and 10.5% respectively. These words serve to pull the overall /k/-deletion up. Second, nine of the ten most common words with /g/ have the phoneme in initial position in a stressed syllable, an unlikely target for deletion. The ten most common /g/ words, in descending order, are get, go, got, gonna, going, good, big, getting, guess, and goes. These lexical facts serve to decrease the overall rate of /g/-deletion.
Table 4.3: Deletion rates (%) for each phoneme in the Buckeye corpus.

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>/t/</th>
<th>/v/</th>
<th>/θ/</th>
<th>/ð/</th>
<th>/s/</th>
<th>/z/</th>
<th>/ʃ/</th>
<th>/ʒ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>3.2</td>
<td>25.8</td>
<td>15.4</td>
<td>31.4</td>
<td>5.4</td>
<td>22.1</td>
<td>8.2</td>
<td>19.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>/p/</th>
<th>/b/</th>
<th>/t/</th>
<th>/d/</th>
<th>/tʃ/</th>
<th>/dʒ/</th>
<th>/k/</th>
<th>/g/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>4.4</td>
<td>11.3</td>
<td>25.9</td>
<td>50.9</td>
<td>13.7</td>
<td>16.3</td>
<td>6.1</td>
<td>4.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>/w/</th>
<th>/l/</th>
<th>/y/</th>
<th>/m/</th>
<th>/n/</th>
<th>/ŋ/</th>
<th>/h/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>7.9</td>
<td>23.4</td>
<td>10.5</td>
<td>11.3</td>
<td>3.8</td>
<td>12.2</td>
<td>6.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>/æ/</th>
<th>/e/</th>
<th>/ei/</th>
<th>/i/</th>
<th>/ə/</th>
<th>/ai/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>8.4</td>
<td>6.6</td>
<td>2.4</td>
<td>10.1</td>
<td>2.6</td>
<td>11.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>/ɑ/</th>
<th>/ou/</th>
<th>/u/</th>
<th>/ʊ/</th>
<th>/ɔ/</th>
<th>/au/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>2.0</td>
<td>6.4</td>
<td>6.6</td>
<td>5.1</td>
<td>2.5</td>
<td>0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.489</td>
<td>-27.108</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Number of phonemes</td>
<td>0.379</td>
<td>23.010</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Log frequency</td>
<td>0.019</td>
<td>0.473</td>
<td>.636</td>
</tr>
<tr>
<td>Number of neighbors</td>
<td>-0.156</td>
<td>-8.407</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Log mentions</td>
<td>-0.017</td>
<td>-1.063</td>
<td>.288</td>
</tr>
<tr>
<td>Phon. × Freq.</td>
<td>0.018</td>
<td>2.799</td>
<td>.005**</td>
</tr>
<tr>
<td>Phon. × Neigh.</td>
<td>0.033</td>
<td>6.589</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Freq. × Neigh.</td>
<td>0.009</td>
<td>4.621</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Phon. × Ment.</td>
<td>0.010</td>
<td>2.617</td>
<td>.009**</td>
</tr>
<tr>
<td>Neigh. × Ment.</td>
<td>0.005</td>
<td>2.189</td>
<td>.029*</td>
</tr>
<tr>
<td>Phon. × Neigh. × Ment.</td>
<td>-0.002</td>
<td>-3.218</td>
<td>.001**</td>
</tr>
</tbody>
</table>

Table 4.4: Output of Poisson regression model predicting number of phoneme deletions per word in the Buckeye corpus.

shorter words. Words with more phonemes are also able to delete phonemes with less likelihood of yielding an ambiguous phoneme string, since neighborhood density is correlated with segmental word length. In line with this reasoning, an effect of neighborhood density was also observed, such that words with more neighbors tended to
Figure 4.6: Mean number of deletions per word as a function of lexical frequency, split by neighborhood density. The different panels show trends in words of differing numbers of phonemes.

have fewer deletions than words with fewer neighbors. Again, this effect is amenable to a similar explanation in terms of ambiguity avoidance: the deletion of a random phoneme from a word with many neighbors is more likely to yield an ambiguous phoneme string than is a random deletion from a word with few neighbors.

Several significant interactions were also observed. Two of these involved frequency: one with number of phonemes, and another with neighborhood density. More frequent words tended to have more deletions than less frequent words, but this trend was diminished for words with few phonemes and for words with many neighbors.
Figure 4.7: Mean number of deletions per word as a function of number of mentions, split by neighborhood density. The different panels show trends in words of differing numbers of phonemes.

These interactions are visualized in Figure 4.6. Note that the bottom panels, representing words with 5 or more phonemes, do not have a curve for words in dense neighborhoods since there were no few words with 5 or more phonemes with more than 10 neighbors. Three other interactions were observed: a two-way interaction between number of phonemes and neighborhood density; a two-way interaction between number of mentions and neighborhood density; and a three-way interaction between number of phonemes, density, and mention. Here, we see a similar trend to the frequency interactions. Words that had been mentioned more tended to have more deletions than words which had been mentioned less; however, this effect was diminished in words
with few phonemes, and words with many neighbors. The three-way interaction suggests that words with many mentions, neighbors, and phonemes did not have quite as many deletions as a na"ive model would predict, suggesting that these effects are not purely linear or additive. This conclusion is supported by general knowledge of lexical structure, which suggests that neighborhood density and number of phonemes tend to be correlated and non-independent. These interactions are visualized in Figure 4.7.

The general trend in these data is that frequency and mention effects exist, but that they are inhibited in words with few phonemes, or words in dense neighborhoods. This inhibition can straightforwardly be interpreted in terms of a communicative effect, whereby talkers avoid rendering words confusable.

4.5.4 Discussion

This analysis of phoneme deletion in the Buckeye corpus has revealed some expected patterns. Words that are more frequent and words in sparser neighborhoods had more phonemes deleted than words that are less frequent and words in denser neighborhoods. These effects were mediated to some extent by the number of phonemes in the word, as depicted in the separate panels of Figures 4.6 and 4.7. Additionally, coronals were more likely to be deleted than non-coronals, and consonants were more likely to be deleted than vowels. Second and subsequent mentions of words had more deletions than first mentions, although this effect was somewhat constrained by neighborhood density.

The lexical factors of frequency and neighborhood density can be seen to be in a complex relationship in the Buckeye corpus, as in the other two corpora examined. Additionally, neighborhood density was observed to influence the effects of the contextual factor of mention in a limited way. These complex interactions between these
effects again speak to the disunity of reduction. There is no single process operating here to control phonetic output, but rather a controlled anarchy of interfering lexical and contextual factors with competing interests.

In experiments 7 and 8, neighborhood density was observed to correlate negatively with vowel duration. That is, words with more neighbors had shorter vowels than words with fewer neighbors. In the analysis of deletion in the Buckeye corpus, however, words with more neighbors were observed to have fewer segment deletions than words with fewer neighbors. Although the data analyzed in experiments 7 and 8 were from read speech, not spontaneous speech, and are thus not easily comparable to the Buckeye corpus, this result presents an interesting avenue for future research. Perhaps segment deletion and general temporal reduction are implemented differently in the production system. Contradictory results for effects of neighborhood density on production have been observed before, but not usually within the same study (e.g. Gahl, 2015; Gahl et al., 2012; Munson and Solomon, 2004; Peramunage et al., 2011; Scarborough, 2010; Wright, 2004).

Interpretations of these analyses must be tempered with some caveats. As with all studies involving frequency and neighborhood density, the effects of these factors cannot easily be separated from those of segmental content, and this concern is especially true in spontaneous speech like the Buckeye corpus. Syllable structure was another variable not controlled for, since it is not marked in the corpus, and likely would have yielded significant effects. For instance, deletion in coda position is generally more likely than in onset position, and open syllables are more likely to delete than closed syllables (Burchfield and Bradlow, 2014). Positional variance in general is another issue with both the current analyses and other contemporary research—for instance, to what extent is a /t/ in word-initial position the same ontological entity as a /t/ in word-final position (see Fricke et al., 2015)? Positional neutralization in
particular causes problems for these kinds of analysis. For example, the voicing contrast in English stops is neutralized in initial position after /s/. The word speech could just as easily be transcribed /sbitʃ/ as it could /spitʃ/. When considering that second phoneme, do we consider it as voiced or voiceless in our models? The decision could have far-reaching consequences—for instance, the word peach is a phonological neighbor of speech only if we choose the latter analysis over the former. Such differences could significantly alter our models of lexical neighborhood structure.

Indeed, the ability to model effects of lexical structure is hampered by an insistence that phonological analyses must be unique and that phonological inventories must be exhaustive. It has been known for several decades that phonological systems can have multiple non-unique potential analyses (Chao, 1934). The ‘non-uniqueness problem’, as it was termed, stirred a brief flurry of research in the 60s and 70s (Hyman, 1970; Malone, 1970; Rubach, 1978; Schane, 1968), particularly concerning whether or not non-unique solutions ought to be permitted in a generative theory phonology. Ultimately, however, the issue has been left largely untouched. More recently, Ladd (2014) has called into question the assumption that phonological inventories are exhaustive—that is, that a listing of a language’s phonemes is necessarily complete. Most languages have some ‘marginal’ phonological contrasts which are intermediate between full contrast and full allophony (Hall, 2013). Similarly, ‘marginal phonemes’ can be said to exist: the fricative /x/ exists in English for an educated pronunciation of Bach; Modern Standard Arabic /l/ is restricted to Allah and a handful of loanwords; as an extreme example, the English expression of disapproval tsk is commonly pronounced as a

---

17 Indeed, the first phoneme is also neutralized, making /zbitʃ/ a possible transcription. In fact, for varieties of English where /s/ and /ʃ/ do not contrast before stops in word-initial position (this contrast was historically introduced into American English via loanwords from Yiddish such as shtik), the transcriptions /ʃpitʃ/ and /ʒbitʃ/ are possibilities.

18 Or tut.
dental click /l/. While traditional linguistics has assumed that ‘being in the phonology’ or even ‘being language’ is a binary relation, these relations may turn out to be more gradient than originally thought. Likewise, the quest for complete systematicity within grammar could prove quixotic. In discussing how to look to synchronic data for evidence for or against an analysis of an indeterminate data set, Malone (1970, p332) averred:

if the synchronic language doesn't provide an answer, the linguist is overstepping his domain in trying to do so. More seriously, he is in jeopardy of chasing a will-o’-the-wisp non-existent to native speakers, a linguistic non-entity. ... There is no reason in the world why a language should not contain pockets of systematic indeterminacy in one or several of its organizational levels; indeed, there is a great deal of long-standing evidence for just such indeterminacy in linguistic change.

Consequently, since the definition of neighborhood density used in the present analysis relies upon an exhaustively and uniquely defined phonology as the basis for the lexicon, the present understanding of density effects could be hopelessly naïve.

Further, the contextual effect of mention considered in these analyses was relatively coarse-grained, simply defined as the number of times a word had been said. However, second mentions are not necessarily reduced if the word has a different referent than the first mention (Bard et al., 1989), and first mentions themselves can be reduced if the referent is contextually salient (Bard and Anderson, 1994). Additionally, the structure of a discourse or narrative can influence mention reduction processes (Fowler et al., 1997; Vajrabhaya and Kapatsinski, 2011). A deeper analysis of the role of mention in the Buckeye corpus, as well as the Clopper corpus, is left to future work (see also Beaver et al., 2007).
4.6 Theoretical ramifications and conclusions

The three experiments presented in this chapter have all suggested that predictability-based reduction is not necessarily a single concept or process. Effects of frequency, neighborhood density, semantic predictability, and discourse mention were observed to have diverse and variable acoustic and phonological consequences. This conclusion was aided by the use of additive models, which allow easy detection and inspection of non-linearities among fixed effects.

Phonetic reduction is clearly a complex phenomenon, and the elements of lexical and contextual predictability which affect it are not fully or even partially understood. Reductionist approaches to these processes can only partially account for the phenomena; for more adequate empirical coverage, more complex hybrid models will have to be specified.

4.6.1 Toward a new measure of neighborhood density

In section 4.5.4, it was noted that analytic decisions, such as the representation of neutralization and syllable position, can drastically alter measures of neighborhood density. Additionally, as outlined in section 1.1.1.2, the standard method of calculation of neighborhood density makes the assumption that all phonemes are equally similar. This method also assumes that neighborhood is a binary relationship—a word is either in or not in another word’s neighborhood. These assumptions do not accord with our current understanding of lexical processing.

Instead, by using a continuous measure of similarity, it may be possible to refine our understanding of what words are similar; such a measure may also lead to a continuous notion of neighborhood. Rather than a word being “similar enough” to be considered a neighbor, every word would instead have a real-numbered similarity
distance from every other word. Such an approach would also allow for subtler distinctions than can currently be made. For example, under the current “edit distance” method, *pat* and *rat* are both equal neighbors of *cat*, while *tack* and *waffle* are equal non-neighbors. However, */p/* is more similar to */k/* than */r/* is, both in terms of articulation and acoustics; we therefore expect *pat* to be closer to *cat* than *rat* is. Similarly, *tack*, being a monosyllabic word with voiceless stops, is more similar to *cat* than is the word *waffle*, which has two syllables and no voiceless stops. A continuous measure would likely derive a continuum of similarity, such that *pat > rat > tack > waffle*, relative to *cat*.

The creation of such a measure would require a database of phonological and psychoacoustic similarity judgements between phonemes, or, preferably, between pairs of phonemes (diphones). A diphone analysis is particularly attractive since it would help control for positional effects, such as the difference between */t/* before and after a vowel. The details and implementation of such a measure are left to future work.

4.6.2 Toward a new understanding of reduction

The goal of this chapter was to examine the conceptual unity of reduction. As discussed in the introduction, the major theories of predictability-based phonetic reduction are reductionist: they attribute reduction to a single factor, such as listener-orientation, talker-orientation, or exemplar dynamics. Hybrid models like that of Watson (2010) propose that there are multiple sources of reduction. One possible approach to a hybrid model is that lexical factors, such as frequency and neighborhood density, form a single coherent factor which influences reduction, while contextual factors, such as semantic predictability and discourse mention, form a second, independent factor. This specific understanding of a hybrid model was tested in this chapter.
Experiments 7 and 8 both examined the effects of lexical factors on reduction, specifically the effects of frequency and neighborhood density on vowel duration and dispersion. The use of GAMMs allowed for relatively simple visualization of complex, non-linear interactions between these factors. The results of these models established that the effects of frequency and neighborhood density are not independent of each other. This result is consistent with a hybrid model, but also is consistent with any of the reductionist models too.

The second analysis of the Clopper corpus in experiment 8 considered the contextual factors of mention and semantic predictability. This analysis showed that first and second mentions differed in vowel duration and vowel dispersion, and, crucially, that the magnitude of the difference between the first and second mention was mediated by the difference between the mentions in their semantic predictability (parameterized by trigram probability). These effects were in the expected directions, such that vowels in second mentions were shorter and less disperse than vowels in first mentions, and vowels in predictable words were shorter and less disperse than vowels in unpredictable words. These results do not provide strong evidence for discourse mention and semantic predictability being manifestations of the same underlying source; rather, their plainly additive effects suggests that these two factors are independent. This result is not compatible with the three main reductionist accounts of reduction, but it could fit into an appropriately-specified hybrid model. The dichotomous lexical versus contextual model outlined above, however, will not do.

Finally, experiment 9 considered a different reduction pattern—phoneme deletion in the Buckeye corpus. This analysis investigated the lexical factors of frequency and neighborhood density and the contextual factor of discourse mention. Word length in phonemes was included as a covariate. Confirming the patterns observed in experiments 7 and 8, frequency and density were both involved in complex interactions
with each other and with word length. Mention interacted with neighborhood density, such that phoneme deletions in multiply-mentioned words were more common for low-density words than for high-density words. While this interaction could be interpreted as suggesting the non-independence of density and mention, caution is advised. Principally, neighborhood density in the Buckeye corpus is hard to tease apart from word length in phonemes. As can be seen in the lower panels of Figure 4.7, there are no words with 5 or more phonemes that have more than 10 lexical neighbors. Therefore, it is difficult to draw generalizations from this result. A more conservative interpretation is that the contextual factor of mention is independent of the lexical factors of frequency and density.

Taken together, these experiments have provided support for the idea that phonetic reduction cannot be attributed to a single cause. This result is consistent with the principles of Watson’s (2010) Multiple Source view. One relevant factor is lexical predictability, consisting of (minimally) word frequency and neighborhood density; another factor is semantic predictability; and another factor is discourse mention. These three factors, identified as distinct from each other in their effects on phonetic reduction, also have conceptually distinct sources. Lexical predictability is essentially static and unchanging throughout the course of a dialogue, and knowledge of these statistics is something the talker has stored in some component of their language system. The number of neighbors of a word, for example, does not change depending on the context. Discourse mention is at the other extreme—this factor is intrinsically contextual. With no context, it is extremely difficult to tell whether a word token is first or second mention. The talker thus cannot store this information; it must be computed in real time. Finally, semantic predictability is somewhere between these two

\[19\] Fowler and Housum (1987) reported accuracy rates of 60% in perception task with a forced-choice of ‘first’ or ‘second’.
extremes. On the one hand, it too is contextual—different circumstances will lead to different predictabilities. On the other hand, it involves some elements of stored knowledge. Presumably the talker has some knowledge of trigrams (or a cognitive equivalent) to consult during the speech production process. Another important piece of semantic predictability is real-world knowledge, and being able to predict upcoming words based on what the sentence is about. The stimuli used in Experiments 1a and 1b were good examples of this real-world knowledge, with sentences like *The shepherd watched his flock of sheep*. Here, knowledge of the fact that sheep are a reasonable thing for a shepherd to watch enables the listener to predict upcoming material, independently of their knowledge of lexical and usage statistics. These three distinct reduction factors—lexical predictability, semantic predictability, and discourse mention—are therefore well-grounded in terms of how we understand language processing.

To summarize, this chapter has examined the conceptual unity of reduction. The three major theories of phonetic reduction explain reduction as arising from a single factor. Watson's (2010) Multiple Source view is a hybrid model which instead proposes that reduction is due to several cognitive sources. The results of the experiments described in this chapter suggest that a hybrid approach is useful in describing patterns of reduction. In particular, three independent sources of reduction have been proposed: lexical predictability, as measured through frequency and neighborhood density; semantic predictability, as measured through trigram probability; and discourse mention, as measured through counting repeated words in a conversation. Further refinement of these measures, and more detailed modeling, will provide more insight into the nature of these factors. Contrary to the reductionist accounts of reduction, reduction cannot be reduced to a single factor.
Chapter 5

*Exploring alternatives to listener-orientation:*

*Theory, evidence, and coherence*

In this final chapter, I first present a summary of the thesis, its main findings, arguments, and conclusions. The remainder of the chapter is devoted to discussion of theoretical alternatives to the listener-oriented account, the challenges these theories could face in accounting for the data presented in this dissertation, and speculative approaches to resolving these challenges.

### 5.1 Summary of thesis

#### 5.1.1 The listener-oriented account of phonetic reduction

As outlined in Chapter 1, the listener-oriented account of predictability effects in phonetic reduction holds that talker behavior is a balance between two competing constraints. On the one hand, talkers seek to minimize effort to conserve energy; on the other hand, talkers seek to maximize the intelligibility of their utterances. Words which are predictable from context are more likely to be perceived correctly than unpredictable words, all else being equal. Therefore, reduction is licensed on those words; reduction is inhibited for unpredictable words. Assuming that reduction and conservation of effort are ‘natural’ or basic processes (Ferrer i Cancho and Solé, 2003; Zipf,
1929), the majority of the work to be done by a listener-oriented mechanism is inhibiting reduction when it is inappropriate.

5.1.2 Testing the prediction relating to theory of mind

The experiments in Chapter 2 assessed the predictions of the listener-oriented account as they pertain to theory of mind (ToM). Namely, for talkers to have competent productions and inhibit reduction in contextually-appropriate places, it is necessary that they are able to create, maintain, and update a model of their interlocutor's mental state. This ability requires a well-developed theory of mind. Therefore, a strong interpretation of the listener-oriented account predicts that extent and consistency of predictability-based reduction should positively correlate with ToM ability.

Experiments 1a and 1b tested this prediction in the domains of reduction due to semantic predictability and second mention. No relationship between second mention reduction and individual differences in ToM scores was observed. However, a negative correlation was observed between ToM ability and semantic predictability-based reduction, such that talkers with poorer theory of mind had larger acoustic differences between predictable and unpredictable words than talkers with better theory of mind. This relationship is in the opposite direction to that predicted by the listener-oriented account.

Experiment 2 investigated the effects of lexical neighborhood structure on the VOT of word-initial voiceless stops, and its relationship with ToM. No effects were observed. It is not clear how to interpret this failure to replicate this relatively well-established effect (Baese-Berk and Goldrick, 2009), and therefore no conclusions about the listener-oriented account can be drawn.
Experiment 3 investigated the role of frequency and neighborhood density in mediating word and vowel production, and how these effects are related to ToM. The results demonstrated an expected effect of frequency on word duration, such that more frequent words had shorter durations than less frequent words. However, a 3-way interaction between ToM skill, frequency, and neighborhood density was observed, such that talkers with a poor theory of mind did not exhibit a frequency effect for words in dense neighborhoods. This finding is somewhat consistent with the prediction of the listener-oriented account, in that talkers with poor ToM were less influenced by frequency in their productions than talkers with better ToM. The listener-oriented account predicts that this effect should be across the board, for all words, but the observed results demonstrated that the effect was restricted to words in dense lexical neighborhoods.

Taken together, these results can be seen as very weak evidence for the predictions of the listener-oriented account, and as much stronger evidence against it. Notably, the effects observed in experiments 1a and 1b were the opposite of those predicted. This conclusion cannot necessarily be interpreted as evidence in favor of alternative accounts, however, since it is similarly unclear how they could explain the observed effects (see Sections 5.2 and 5.3).

5.1.3 Testing the assumption of a perceptual benefit

Chapter 3 tested the foundational listener-oriented assumption that unreduced speech confers a perceptual benefit relative to reduced speech. The listener-oriented account
assumes that talker behavior strikes a balance between conservation of effort (hypoarticulation) and intelligibility (hyperarticulation). Implicit in this account is the assumption that hyperarticulated words are necessarily more intelligible and otherwise beneficial to the perceptual system, relative to hypoarticulated words. While this assumption appears to be obviously correct for extreme cases of reduction and hyperarticulation, it is not clear if it is true for smaller differences along the hypo-hyper continuum, nor is it clear if this purported effect solely influences intelligibility, or if other perceptual processes are affected too.

Although the perception of reduced speech has been examined before, with the general conclusion that reduced speech is harder to process (Pitt, 2009; Ranbom et al., 2009; Tucker and Warner, 2007), the experiments in Chapter 3 differed from the previous literature in a number of respects. One of the larger differences was the nature of the stimuli. Many studies of the perception of reduced speech examine categorical reductions such as /t/-flapping (Tucker and Warner, 2007) or use speech spoken by a trained phonetician (Ranbom and Connine, 2007), rather than natural speech from naïve talkers (the work of Mirjam Ernestus is a notable exception in this regard; see Ernestus et al., 2002, inter alia). The stimuli used in the experiments in Chapter 3 were taken from the natural read speech collected in experiment 1a. The reductions typified in this speech were small gradient durational distinctions, rather than the massive (Ernestus et al., 2002) or categorical (Tucker and Warner, 2007) reductions examined in previous studies. Another novel aspect was the investigation of individual differences in the ToM of the listeners. While some studies of speech perception have examined individual differences in cognition, particularly within research on aging (see Akeroyd, 2008, for review), theory of mind remains almost wholly understudied (cf. some exploration of autistic traits: Stewart and Ota, 2008; Yu, 2010, 2013). The final novel aspect
of these experiments was their examination of multiple levels of processing—from subjective judgements of clarity to semantic processing. Each of these levels have been examined by individual studies in the past (e.g. Ferguson and Kerr, 2009; McLennan and Luce, 2005), but I am unaware of any single study of several levels which uses a set of stimuli drawn from the same source throughout. The use of the same set of stimuli is beneficial since it allows for direct comparison between experimental paradigms.

Experiment 4 used a subjective judgement task to assess the perceived clarity of the stimuli, and also a speech perception in noise task to assess the intelligibility of these words. Experiment 5 was an auditory lexical decision task, examining how quickly and accurately listeners were able to make lexicality judgements. Finally, experiment 6 involved the stimulus words embedded in semantically biased or neutral sentences, and used a semantic acceptability task to determine the speed and accuracy with which listeners were able to judge the semantic acceptability of the stimulus sentences. In all experiments, phonetically reduced (shorter) words led to slower and less accurate responses than unreduced words. This finding vindicates the foundational assumption of the listener-oriented account.

However, individual differences were observed which are difficult to square with the listener-oriented account. In the lexical decision task, listeners with poorer ToM skills experienced a larger slowdown in reaction time for stimuli of a short duration than listeners with better ToM skills. That is, the low ToM listeners were more adversely affected by the stimuli with limited phonetic content than the high ToM listeners. This result was interpreted as being due to the low ToM listeners’ heightened reliance on acoustic rather than top-down cues in their parsing. While the high ToM listeners may have relied on knowledge of lexical frequency and phonotactic probability to assist them in their decision, the low ToM listeners instead paid relatively more attention to the acoustics, which led to longer reaction times for phonetically reduced stimuli.
Additionally, in the sentence acceptability task, there was evidence that low-ToM listeners were again more sensitive to fine acoustic detail than high-ToM listeners. Their response accuracy was influenced by an interaction between stimulus duration and stimulus production context (essentially an acoustic factor, the precise correlates of which are unknown), while the responses of the high-ToM listeners were not affected by these acoustic factors. Again, this finding speaks to a correlation between ToM and acoustic cue attention. This finding is consistent with work on the autism phenotype suggesting heightened attention to acoustic information over linguistic or contextual information (Imaizumi et al., 2009; Järvinen-Pasley et al., 2008; Mottron et al., 2006).

5.1.4 Reconsidering the concept of reduction

Finally, chapter 4 used corpus analysis to critically examine ‘phonetic reduction’ as a coherent concept. In addition to mixed effect regression models, which are fast becoming the de facto standard for inference in linguistic studies, this chapter utilized generalized additive mixed modeling (GAMM) to probe non-linearities among covariates.

Experiment 7 was a GAMM analysis of the data from experiments 1a, 1b, and 3. The results revealed significant effects of frequency and density, such that more frequent words and words with more neighbors, underwent more reduction than less frequent words and words with fewer neighbors. Notably, these effects were not strictly linearly related to each other, supporting the idea that underlyingly they represent the same factor. Experiment 8 repeated the analysis of experiment 7 on the Clopper corpus and confirmed the lack of independence between neighborhood density and frequency. Additionally, this experiment also investigated the effects of second mention and trigram probability on reduction in this corpus. Effects of both were found, but in a strictly linear relationship to each other, suggesting that these factors have different
cognitive sources. Finally, experiment 9 considered phoneme deletions in the Buckeye corpus as a function of lexical frequency, neighborhood density, and discourse mention. Again, frequency and density patterned together, confirming the conclusions of experiments 7 and 8. Discourse mention, however, was generally independent of these factors, confirming the conclusion of experiment 8 that it is a separate factor.

Taken together, these diverse findings illustrate that the acoustic and phonological consequences of reduction are various and that the interactions between the causes of reduction are extremely complex. Specifically, lexical factors such as frequency and neighborhood density, discourse factors such as mention, and predictability factors such as trigram probability are all independent factors with different contributions to reduction.

5.1.5 Original contribution to knowledge

The major contributions of this thesis to our knowledge of the world concern three areas. The first of these concerns individual differences in cognitive style and their relation to language use. Chapters 2 and 3 both revealed non-trivial and somewhat unexpected roles of individual variation in theory of mind in mediating aspects of language production and language perception. These traits in particular, and individual differences in general, are understudied within linguistics, especially outside of a clinical or pedagogical setting. Linguists are now used to thinking about how social or regional variation may affect language and use of language. Perhaps now it is time to consider how cognitive and personality variation may also affect linguistic processes.

The second major contribution of this thesis concerns phonetic reduction. Far from being a unitary phenomenon reducible to a single process or constraint, the nine experiments in this work have all demonstrated the remarkable diversity of reduction,
in terms of origins, effects on acoustics, and effects on perception. While some researchers may regard this conclusion as self-evident (see e.g. Clopper and Turnbull, 2015), the sheer number of research articles which generalize their results from one particular instance of reduction to ‘reduction as a whole’ speaks to the fact that reduction’s disunity is not self-evident to all.

Finally, the listener-oriented account, the main topic of this thesis, constitutes the third major contribution—particularly, the results herein suggest that the listener-oriented account is relatively weakly supported. Although its basic assumptions were vindicated in Chapter 3, none of its main predictions were borne out unambiguously, and there exist a host of phenomena discussed in this thesis which the theory cannot easily account for. With that being said, this argumentation against the listener-oriented account should not be taken as evidence in favor of other accounts. As we shall see, alternatives to the listener-oriented account have problems of their own, even in accounting for the data presented in this thesis. To these alternatives we now turn.

5.2 Talker-oriented models

A common mechanism in many talker-oriented models is that less frequent words are harder to access, and that this difficulty causes a slowdown in the speech production process, which in turn affects word and vowel duration (Bell et al., 2009, 2003; Fox Tree and Clark, 1997). This account finds support in other domains, for example in findings that disfluent utterances are often articulated more slowly and more fully than fluent utterances (Munson, 2007), that disfluent utterances are more likely to refer to discourse-new items than to discourse-given items (Arnold and Tanenhaus, 2011), and that less frequent words are accessed slower in perception (Balota and Chumbley,
These pieces of evidence suggest that words which are harder to access by virtue of their frequency or low salience (in the case of discourse) are more likely to cause a slowdown in production, and thus disfluency, than easy words. Bell et al. (2009) specifically proposed a mechanism which “modulates the pace of articulation” (p105) when phonological encoding is slowed or delayed. This mechanism keeps the physical articulation and the psychological lexical stream synchronized—if the words cannot be accessed quickly enough, the mouth slows down to compensate. Consequently, the phonological encoding of low frequency words is slowed, “but not so severely to require disfluent adaptations” (p106), and speech articulation is slowed as a result.

This model claims that the duration reduction on high frequency words (or the duration increase on low frequency words) is entirely due to talker-internal processes. The proposed articulation modulation mechanism means that speech articulation is constrained by speed of phonological encoding. Since language users are able to speak quickly or slowly at will, there are a number of interesting predictions made by this model relating to speech rate. First, we assume the null hypothesis that consciously-controlled speech rate does not influence lexical processing, and that lexical processing does not influence consciously-controlled speech rate. This assumption means that lexical processing advances at the same speed regardless of how fast or slow a talker is trying to speak. Therefore, in slow speech, there is more time for phonological encoding to occur; it follows that in slow speech comparatively smaller durational increases (i.e. smaller levels of slowdown) for infrequent words will be observed, due to the ‘lexical buffer’ of speaking slowly. Conversely, in fast speech, comparatively larger durational increases (i.e. greater levels of slowdown) are expected to be observed, due to the speed of articulation relative to the phonological encoding. Since this model claims that predictability effects are caused by delays in lexical processing, it is conceivable that if lexical processing is artificially delayed, then the predictability effects would be
larger (i.e. more slowdown). Lexical processing is a cognitive task and can therefore be interfered with by taxing cognitive resources through attentional shifting or memory load. If there are fewer resources available to devote to phonological encoding, durational increases are expected to be larger.

Bell et al.’s (2009) model predicts that in faster speech, predictability effects are stronger, and that in instances of high cognitive load, predictability effects are stronger. Evidence that speech rate and cognitive load have an effect on the size of frequency-mediated temporal adjustments would provide evidence in favor of this model of predictability effects. Alternatively, a listener-oriented model does not make any prediction with regards to speech rate. Since a listener-oriented model relies on online modeling of the interlocutor, it makes a prediction for cognitive load—in conditions where cognitive resources are taxed, predictability effects are expected to be applied more varyingly and inconsistently than in a non-distracted condition.

These predictions have yet to be tested, although an analysis of the Clopper corpus can provide a preliminary test of the speech rate prediction. A linear mixed effects regression model was constructed to predict vowel duration, with log word frequency, speech style (clear versus plain), and their interaction as fixed effects. Random intercepts of talker and word, with random slopes of log frequency and style, were also included. The articulation modulation mechanism predicts that in faster speech (in this case, plain speech), frequency effects are larger (more difference between high- and low-predictability words) than in slower (clear) speech. That is, an interaction between frequency and style is predicted. The model revealed expected significant effects of frequency: more frequent words had shorter vowels than less frequent words ($\beta = -5.255, t = -1.994$); and style: plain speech had shorter vowels than clear speech ($\beta = -20.152, t = -5.943$). A significant interaction was also observed ($\beta = 2.4560, t = 3.340$), but in the opposite direction to that predicted by
Figure 5.1: Vowel duration in the Clopper corpus as a function of log word frequency, split by speech style. Linear trend shown.

the articulation modulation mechanism: frequency effects were in fact larger in clear speech than in plain speech, as shown in Figure 5.1.

This result can only be taken as preliminary: many factors influence vowel duration beyond just frequency and speech style, and the Clopper corpus was not designed to test the articulation modulation mechanism. Additionally, inferring conclusions about the articulation modulation hypothesis from the results of this quick test is hampered by the conflation of clear speech with slow speech, and of plain speech with fast speech, since the dimensions of speed and of clarity are to some extent orthogonal and independently controlled (Amano-Kusumoto et al., 2014; Krause and Braida, 2002, 2004). Nevertheless, this preliminary test suggests that the predictions of the articulation modulation hypothesis merit further investigation.
A careful study involving manipulation of speech rate would help to test this prediction. The majority of prior literature involving speech rate manipulates this variable by explicitly asking the participant to speak quickly or slowly (e.g. Mefferd and Green, 2010). Although this approach yields more natural speech than, say, aligning words with the beat of a metronome, there is no guarantee that there is inter-talker or even intra-talker consistency in the realization of these different speech styles. That is to say, sometimes ‘fast’ speech may be faster than other times, and sometimes it will be slower. In light of these problems, Krause and Braid (2002) used a novel procedure to establish and regulate comfortable speech rates for their talkers. Talkers heard a metronome ticking at the desired rate, and were instructed to fit each sentence between two consecutive metronome clicks. Talkers were able to take as much time as they desired between sentences. This method kept the rate of speech consistent across trials, while still allowing the talker the freedom to enhance or reduce particular syllables or words as they see fit. Downsides, however, include the fact that it requires all stimulus sentences to be of approximately the same length; and that in speaking in time with a metronome, talkers are carrying out significantly more perceptual monitoring of their own speech than we can reasonably assume them to do in natural linguistic interactions outside of the laboratory.

Another approach to testing the articulatory modulation hypothesis is through cognitive load. Since lexical processing is a cognitive task, it can be interfered with by taxing cognitive resources through attentional shifting or memory load. According to the articulatory modulation hypothesis, if there are fewer resources available to devote to phonological encoding, durational increases due to predictability factors should be larger. A method that could be applied to this study is a variant of Haarmann et al.’s (2003) **conceptual span** task, which measures semantic short-term memory ability. In this task, the participant is presented with orthographic representations of nine nouns,
one every second, for example “lamp, pear, tiger, apple, grape, elephant, horse, fax, phone.” Each of the nouns belongs to one of three conceptual categories. Next, the participant is presented with a short sentence, which they will read aloud. Immediately after the sentence has been read, the participant is prompted to recall aloud the three members of a named category, in any order: for example, if the prompt is “fruit” the correct response is “pear, apple, grape” in any order. This procedure forces the participant to concentrate on remembering the words they have read, and therefore devote less attention to the production portion of the task. Another memory-based distractor task that could be used in this type of investigation involves presentation of phonotactically legal nonwords (e.g. ‘vong’, ‘treane’). After a sentence reading task, the participant is prompted to freely recall as many of the nonwords as possible (Mattys et al., 2009).

The results of such investigation of the predictions of the articulation modulation could lend credence to the talker-oriented account of predictability effects. Alternatively, if the predictions are not borne out, the case for the talker-oriented account is weakened.

5.3 Passive evolutionary accounts

Another alternative to listener-based modeling is the ‘evolutionary’ accounts advocated by Blevins (2004), Pierrehumbert (2001a, 2002, 2003a), Silverman (2012), and others. This account is related to the listener-based accounts in that it is based on the principles of maximizing intelligibility, but there are no active forces or constraints promoting this behavior. Rather, it is an emergent property of language as a communication system. This explanation has a number of appealing features, some of which derive from taking a ‘dynamic’ or ‘complex systems’ perspective on the problem.
5.3.1 Exemplar dynamics

In an exemplar model (see Johnson, 1997; Pierrehumbert, 2001a), categories are stored as clouds of remembered tokens. For our purposes, the categories we will be considering are words, and each token is a perceived utterance of that category. These exemplar clouds are organized such that similar tokens are closer together, while dissimilar tokens are farther apart. The cloud is a highly multidimensional space with several psychoacoustically relevant parameters. Also recorded for each word token are contextual properties of the token, such as “formal setting” or “child’s speech”. Over time, it is assumed that exemplar representations decay—that is, their activation decreases.

Two mechanisms in this model conspire to give rise to predictability effects. The first is the mechanism by which new tokens enter the system. As new tokens are perceived, they are classified into existing categories according to their likelihood of having originated from that category given their perceptual distance from existing category members. Crucially, if a token is misperceived, it is discarded and not recorded in any category. The second mechanism is how speech production unfolds in this system. In production, a token is selected at random from the exemplar cloud and used as a phonetic target. These two mechanisms conspire to cause predictability effects. To illustrate this ‘conspiracy’, we will consider word frequency (high or low) and phonetic reduction (reduced or unreduced). All else being equal, a high-frequency word token

---

1 This characterization is necessarily simplified and glosses over many details, several of which are also left underspecified by Johnson (1997) and Pierrehumbert (2001a). There may well be complex factors which influence the choice of a phonetic target, relating to activation level, speech style, social situation, and a myriad of others. Similarly, it is not specified if the exemplar tokens themselves contain motor programs, gestural scores, or similar articulatory details, or if they are purely (psycho)acoustic and motor specification is computed online. This question also raises the interesting issue of ‘normalization’ and the abstract nature of speech sounds. Despite the fact that a child’s production of [u] is, in terms of formant frequencies, very similar to an adult man’s production of [æ], these sounds are rarely misclassified or mispronounced.
(H) has more likelihood of being correctly perceived than a low-frequency word token (L), due to its inherent property of frequency. Similarly, all else being equal, an unreduced word token (U) is more likely to be perceived correctly than a reduced word token (R). Together, these relationships yield the following hierarchy of perceptibility: $\text{HU} > \text{HR} \approx \text{LU} > \text{LR}$. Thus, the tokens that are the least likely to be correctly perceived and stored in the exemplar cloud are the reduced low-frequency tokens. Due to these tokens not being as commonly stored in the exemplar cloud, they are also less common in production, relative to the other tokens. This situation quickly becomes a self-reinforcing feedback loop as communication occurs between people. Since LR tokens are now less commonly produced, their chances of being stored in the exemplar cloud becomes even smaller, until LR tokens have disappeared from the system entirely, leaving only HU, HR, and LU tokens. This pattern thus yields the standard frequency effect—high frequency words are phonetically reduced, but low-frequency words are not.

This reasoning can be extended to gradient measures of frequency and reduction with relative ease, where the model yields a continuum of reduction size relative to frequency. These mechanisms can also account for the effects of other lexical factors such as neighborhood density, by assuming that words in sparse neighborhoods are more easily perceived, all else being equal, than words in dense neighborhoods. The same discarding of misperceived tokens, and subsequent random selection from the exemplar cloud in production, produces effects whereby words in dense neighborhoods are hyperarticulated relative to words in sparse neighborhoods.

It is less clear, however, that an exemplar approach is able to account for predictability effects relating to context. Since these effects are by definition contextual and changing in real-time, the long-term nature of exemplar dynamics do not have an
appropriate mechanism to alter productions ‘on the fly’. It is possible that, given sufficiently enriched exemplar representations which are able to take into account whether a token was uttered in a low- or high-predictability context, context-appropriate tokens could be chosen by a speech production mechanism which is sensitive to the predictability context. However, such an approach would require a large number of tokens before systematic effects could be observed, and would not easily generalize to novel words.

The exemplar dynamics approach is radically different from both the listener- and talker-oriented perspectives in that there is no specific constraint or mechanism proposed for predictability effects. Instead, these patterns of behavior emerge out of the interactions of simple components (word tokens). Such an approach is similar to the study of complex systems and emergent patterns, which has some conceptual advantages over other explanations.

5.3.2 Complex systems and emergence

An example of a complex system interaction is the organization of sediment bedforms such as sand dunes or silt in riverbeds. Bedforms tend to orient themselves either parallel or perpendicular to the prevailing flow (either of wind or water) in an area. Rubin and Hunter (1987) proposed that bedforms align in such a way that maximizes sediment transport. This theory was confirmed by a longitudinal experimental study which tracked the formation of novel dunes and prevailing wind conditions in the Tengger Desert in Inner Mongolia (Ping et al., 2014). The precise technical details of these studies is not immediately relevant. The important generalization is that sand dunes provide an example of a fully physically deterministic complex system of particles which organize themselves into a large-scale pattern. This organization follows a
principle of maximization of sediment transportation, which can be solved in different ways (parallel or perpendicular to prevailing flow). It is possible to anthropomorphize the dune itself, and to claim that the dune ‘wants to’ maximize sediment transport. It is possible to postulate a constraint or a principle, or even some active cognitive force on the part of the sand dune, which ensures that this maximization takes place. Such postulations are unnecessary, since the explanation of the formation of dunes as a consequence of interactions among fluid dynamics is sufficient.

Similar reasoning can be applied to the evolutionary account of predictability effects. If an explanation in terms of emergent properties of the interactions among exemplar dynamics is sufficient, then any cognitive architecture—be it listener-based or talker-based—is unnecessary. As demonstrated earlier, the exemplar dynamics account is able to explain lexical predictability effects with ease.

Conceptually similar reasoning on the role and importance of such constraints comes from the theory of substance free phonology (Blaho, 2008, chapter 1). Hale and Reiss (2008, section 6.4) presented a thought experiment where a new species of human exists. These humans are identical to existing humans in every way, except that they have a membrane at the base of their oral cavities which can expand under pressure, much like a frog or toad. Their description of this membrane is very similar to the laryngeal air sac possessed by most non-human primates, and hypothesized to aid in call vocalizations (Hewitt et al., 2002; Nishimura et al., 2007). Hale and Reiss (2008) asked the question of whether these modified humans would be able to learn German (or English, or Hawaiian) if placed in the same acquisition environment as a normal human. The natural response is yes, since their cognitive structure is identical and the only difference is in the membrane. Hale and Reiss (2008) pointed out that, due to the

---

2The loss of air sacs in human evolutionary history is incidentally a fascinating case, especially in terms of how this loss may have affected speech (de Boer, 2012; Littauer, 2012), but not relevant to the present discussion.
membrane, the physiological grounding for the typological bias against voiced obstru-
ents in codas has disappeared.\(^3\) Despite this physical difference, these humans are still able to learn languages with different treatments of obstruent voicing (German, En-
glish, and Hawai’ian). Hale and Reiss (2008) used this reasoning to advance their the-
ory that phonetic substance does not belong in phonology, meaning that (for instance) constraints such as *\(\text{V}^{\text{OICED}}\)-\(\text{CODA}\) are not feasibly part of a cognitive system (Odden, 2010). However, I wish to present a slightly different interpretation, albeit one that I imagine Hale and Reiss (2008) would not object to. The results of this thought exper-
iment highlight the fact that typological properties of language can easily be functions of the physical constraints of the space the language is used in. Indeed, biophysics has been claimed to be of central importance in understanding the development of species, even more than natural selection (Akre and Johnsen, 2014; Thompson, 1917). Cross-
linguistic vowel patterns have been suggested to be the result of acoustic properties of the morphology of the vocal tract (Carré, 2009). Extending this reasoning to phonetic reduction and predictability, a compelling argument against the cognitive constraints proposed by the listener- and talker-oriented accounts would be one which is able to account for reduction as an emergent property of the communicative system.

If it can be demonstrated that exemplar dynamics and similar evolutionary mechanisms are sufficient to account for predictability effects, then the evolutionary approach is preferable to the listener- or talker-oriented accounts on the grounds of its simplicity. However, in addition to the empirical issues with contextual predictability effects noted in Section 5.3.1, there are also some conceptual issues with this account.

\(^3\)The specific reasoning behind this claim relies on the idea that voicing coda obstruents is physiologi-
ically difficult (Westbury and Keating, 1986). However, regardless of the specifics of the claim, it is trivial to come up with a new species of humans which differs from existing humans in a single non-cognitive factor which has been implicated in language change and use. The results of the thought experiment are the same.
5.3.3 The evolution of communication and language

The passive evolutionary approach critically relies upon a fitness mechanism. In biological evolution, the fitness of a phenotype determines the likelihood of its being reproduced and passed on to future generations. The fitness mechanism in this case is the survival and successful reproduction of the individual that carries the phenotype. The mechanism rewards reproduction and penalizes death, and the system thus evolves to a state where the former is maximized and the latter minimized. In the exemplar dynamic model sketched above, word tokens are the items which can reproduce and die within a linguistic ecosystem consisting of several individual agents communicating. Successful reproduction of a word token, in this model, is the word’s being selected for production and being perceived correctly. Thus, the word has spread from one individual to another. The death of a word token is the word’s being perceived incorrectly, and thus discarded from a listener’s exemplar cloud. The complete extinction of a word token would take place if all the individuals who possess that token die without reproducing the word token.

In this system, the fitness mechanism is not as obvious as that of biological evolution. The passive evolutionary approach holds that the fitness mechanism is communication. Silverman (2012, p147) wrote that “[s]uccessful speech propagates; unsuccessful speech does not”, proposing that correctly-perceived tokens are stored in a listener’s exemplar cloud, while misheard tokens are discarded. Tupper (2014) has provided the mathematical and conceptual basis for this process of discarding. A minor problem with this account is that it requires listeners to be aware of both what they perceived the token as and what the token actually was. In real life, listeners only have access to the former, not the latter.\footnote{The exception is when errors are explicitly noted as part of the discourse.} However, this problem is only minor since
it is relatively simple to build in new mechanisms by which a listener rejects tokens which are sufficiently vague, quantified by confidence of categorization which can easily be calculated by treating the categorization as a probability estimation (Ashby and Alfonso-Reese, 1995). Tokens can be vague either due to problems with the signal or the context. A problem with the signal could be extreme phonetic reduction or the sound of a passing truck. A problem with the context could be the perception of a syntactically illegal sequence or an incongruous meaning. Sufficiently vague tokens are thus discarded.

In this type of model, unsuccessful communication is penalized. Word tokens which are not understood never enter an exemplar cloud, and thus disappear from the linguistic ecosystem. The system then maximizes word tokens which lead to successful communication, and minimizes words which fail to communicate successfully. This model emphasizes the idea that word tokens have ‘private lives’ of their own as they are propagated (or not) throughout a speech community. That is, the word tokens exist to some extent independently of the talkers and listeners.

However, it should be borne in mind that people, the bearers of exemplar clouds, also have private lives, and that these private lives exist within a complex social ecosystem. That is, just as individual animals interacting within an ecosystem have the larger environment acting asymmetrically upon them, so too do words interacting within a set of exemplar clouds have larger social and physical environments acting upon them. Two examples will suffice to illustrate.

First, it is possible that successful communication is not rewarded. In an extreme case, such as picking a fight with the wrong person, communicating successfully can lead to death, while not communicating at all may lead to survival. Less extreme

---

5By asymmetrical action, I am referring to the fact that, for example, the Chicxulub asteroid acted upon the dinosaurs in such a way to influence their fitness, but the dinosaurs were unable to act upon the Chicxulub asteroid to influence its fitness.
cases are possible where successful communication can lead to negative social consequences, such as ostracism or exile. Note here that successful communication is not necessarily the same thing as truthful communication, and indeed successfully communicating a lie which is later revealed as false may have a net negative effect on the individual fitness of a person. In such a scenario, not having communicated the lie in the first place may have led to better fitness. In both cases, death and exile, the ‘reproduction’ of the word tokens is significantly hampered. Second, it is possible that successful communication does not lead to propagation of the tokens. This scenario can arise when social meaning interferes and blocks or obstructs the usage of particular forms. One path for obstruction is the avoidance of taboo words: some people will never say *fuck*, despite knowing perfectly well what the word means. Another path is the indexing of identity: a Brit would not say *sidewalk*, unless they wished to index a non-British identity. Despite the avoidance of production, the perception and comprehension of the word is unhindered. It can be seen, then, that ‘word fitness’ is not only dependent upon communication, but other factors too.

However these factors do not necessarily invalidate or disprove the mechanisms underlying exemplar dynamics. To be sure, these are factors that must be taken into consideration for a full view of language, but the existence of social factors does not abrogate the fitness mechanisms for word tokens. To use an example from biological evolution, the inability of the dinosaurs to shoot the Chicxulub asteroid out of the sky\(^6\) does not invalidate the role of Darwinian fitness in their evolution. Similarly, the fact that social factors can quash any exemplars that they please does not mean in and of itself that the proposed fitness mechanism underlying their propagation is false. Falsification of this mechanism must proceed from independent evidence.

\(^6\)Rumored fossils of the fabled *Brucewillisaurus* have yet to be confirmed.
There may be such independent evidence causing us to question whether the fitness mechanism postulated for the passive evolutionary account actually exists. The nativist position on language evolution (Berwick et al., 2013) holds that communication is simply a convenient side-effect of language (see also Chomsky, 2002, p144–150). Indeed, the majority of simulation studies of language evolution have not used communicative success as a major factor (see Perfors, 2002, for review). Nevertheless, most researchers in language evolution acknowledge the role of communication (see Christiansen and Kirby, 2003, for review) even if it is not central to their theory of how language evolved.

Dunbar (1996) put forward a theory of language evolution based on grooming. The majority of non-human primates live in groups and engage in social grooming as a way to maintain relationships. In earlier work, Dunbar (1992) had noted a relationship between brain size and group size in primates, with larger brained primates tending to have larger groups. In the evolutionary history of humans, brain size has increased substantially, and it is reasonable to assume that group size has also increased. As group size increases, the time required to adequately groom other group members for the maintenance of social order becomes maladaptive. Dunbar (1996) suggested that vocal behavior evolved to fill this gap, and this behavior laid the foundations for language. It is possible to speak to several individuals at once, but grooming is a one-on-one activity. A survey by McComb and Semple (2005) of forty-two species of primates found strong correlations between group size and size of vocal repertoire, as predicted by Dunbar’s (1996) account. This theory has been extended to a general principle of all animal groups, whereby social complexity drives communicative complexity (Freeberg et al., 2012); evidence in favor of this hypothesis has been observed in a diverse range of animals from chickadees (Freeberg, 2006) to sciurids (Blumstein and Armitage, 1997).
In this theory of language evolution, successful communication of propositional content is not the driving force behind these vocal behaviors. Instead, Dunbar (1996) suggested, language evolved as a mechanism by which relationships can be maintained; phatic communication is primary, meaning is secondary. Experimental support for Dunbar’s (1996) position comes from Fay et al.’s (2008) study of ad hoc communication systems evolved in an experiment with adults. The systems that were evolved with several people interacting—i.e. a group—tended to be better than the systems evolved in purely dyadic interactions. The group communication systems had more efficient information coding, were learned easier, and were more understandable for a naïve participant who was new to the system. This result, and subsequent research (Fay et al., 2010), supports the notion that the development of language critically relied upon group interaction, and additionally that fitness mechanisms could well be related to learnability and least effort rather than communication per se.

5.3.4 Lexical versus contextual factors

The passive evolutionary account presents a coherent story of how predictability effects can emerge and be maintained within a lexicon in a language community. Although there are some potential issues relating to the proposed evolutionary history of the account, these issues are relatively small and rely on an assumed linking hypothesis which explicitly connects phylogeny to contemporary cognition, which is not necessarily justified (see also Hauser et al., 2014).

However, as much as the passive evolutionary account is able to explain phonetic reduction due to lexical factors such as frequency and neighborhood density, it is completely unable to explain phonetic reduction due to contextual factors such as
second mention reduction or semantic predictability. This fact follows from the passive nature of the theory, and the fact that there are no contextual adjustments made during speech production—talkers simply sample from their exemplar clouds without concern for the situation. A reasonable approach to a solution is to therefore supplement or enhance the evolutionary account with some on-line mechanisms, such as listener-modeling or talker-oriented architecture, to account for contextually-motivated reduction.

5.4 Toward a solution: a hybrid model

Although the passive evolutionary account provides a parsimonious account of lexical phenomena, it fails at dealing with contextual phenomena. It may be possible to use this model as a base and add mechanisms as needed, creating a hybrid model like Watson’s (2010) Multiple Sources view. More concretely, we can begin with Pierre-humbert’s (2001a) exemplar dynamics model. As explained earlier, this model is able to account for the lexical effects of frequency and neighborhood density on reduction. Effects of ToM, such as that observed in experiment 3 where participants with poor ToM were less influenced by frequency information than participants with better ToM, can be modeled by less refined or well-established frequency representations for the poor ToM talkers.

To account for contextual effects such as second mention reduction we require a model of the common ground.\textsuperscript{7} Essentially a model of the discourse context, this concept is where the interlocutor keeps track of the question under discussion (Roberts, 2012) and shared knowledge. This common ground approximates a model of the listener, but note that the understanding of common ground being promoted here is a

\textsuperscript{7}Such a model will also assist in accounting for similar discourse patterns such as conceptual pacts (Brennan and Clark, 1996).
purely egocentric one. That is, the common ground model is unique to each talker, and they may match or mismatch the common ground model of their interlocutor (Kecskes, 2007). This common ground model contains merely the information that the interlocutor *assumes* is shared. While these assumptions are often correct, they can be incorrect (Bard et al., 2000; Horton and Keysar, 1996) and the common ground is not necessarily updated correctly or efficiently (Keysar et al., 2003). This last point suggests that updating and maintaining a model of common ground is effortful or requires attentional resources (Schneider et al., 2012b, 2014). The assumption of an egocentric model of common ground which, at best, only approximates the listener, is motivated by the results of this dissertation: notably, the lack of strong evidence linking theory of mind to listener modeling. A richly-represented model of the listener would necessarily vary as a function of individual variation in theory of mind, which was not observed. The proposed common ground model does not require such a theory of mind link, since it is a model of *assumed* shared knowledge. Anecdotally, many people report interactions with others who radically under- or over-estimate the level of shared knowledge, and interactions can suffer as a result (see also Mey, 2008).

Speech production in this model follows the exemplar dynamic approach of selection of a token from the appropriate exemplar cloud. However, prior to this step, the common ground is consulted. Additionally, perception is intimately tied to production, and both of these processes require each other to function effectively. Together with perception information, the common ground informs the selection and implementation of a token from the exemplar cloud. At this stage, modification of the tokens in production is permitted. If the common ground licenses reduction of an element, it is reduced; if consideration of comprehension reveals a difficult contrast, that contrast is enhanced. Again, note here that the talker is essentially consulting their own model of comprehension and common ground, which will result in suboptimal productions if it
does not match that of the listener. Again, such consultation is an effortful process, and talkers can fail to consider comprehension fully.

By tracking, updating, and maintaining the common ground, the talker is able to model the listener in such a way that approximates the predictability effects observed in the literature. More frequent words are reduced relative to less frequent words by ‘default’, since that is part of their exemplar representation. These productions are then modified according to the specifications of the common ground, which in many cases will match well what the listeners knows. Crucially, the implementation of these modifications must be language-specific, to account for acoustic differences in predictability effects between languages (Holliday and Turnbull, 2015), and indeed to account for reduction phenomena in signed languages (Hoetjes et al., 2012). Therefore, this model cannot ‘hard-wire’ a connection from cognition to articulation, but they must be related by general principles.

The model just outlined is necessarily vague, general, and exploratory. However, a somewhat similar model already exists, in some respects. Pickering and Garrod (2013) put forward a theory where language production and comprehension are integrated processes, each relying on each other. In this model, both production and comprehension proceed with continual prediction of upcoming events, comparing the current state to a predicted outcome, and making adjustments as necessary. In this regard the model draws heavily upon control theoretic accounts of, say, skilled limb movement. For example, in reaching to grasp an item, a series of muscle movements must be planned and coordinated. During the movement, biofeedback (somatic and visual) permits changes to the planned action if the predicted outcomes were incorrect (e.g. the object is farther away than previously thought). The reliance on prediction at all stages of the process ties this model nicely to individual differences, such that
skilled predictors should be more adept at speech tasks than unskilled predictors. Indeed, Sinha et al. (2014) have proposed that prediction is one of the main cognitive deficits of autism, a view which could be integrated with the theory of mind results in this dissertation. Pickering and Garrod (2013) support their thesis by noting that listeners tend to predict the speech of their interlocutor (see Pickering and Garrod, 2007, for review), that interlocutors regularly overlap in their speech and jointly negotiate meaning through dialogue (Gregoromichelaki et al., 2011; Knutsen and Le Bigot, 2012, 2014), and that people are remarkably good at speaking in synchrony, also known as “joint speech” (Cummins, 2009). While the details are relatively general, such that concrete predictions are difficult to derive from the model, the basis of production and perception being intimately linked appears to be on solid theoretical ground (but cf. Remez, 2015). Pickering and Garrod’s (2013) model does not contain an explicit account of common ground, but earlier work by the same authors (Pickering and Garrod, 2004) outlined mechanisms by which interlocutors automatically align and repair their representations during the course of a dialogue. This model therefore holds much promise for a unified, hybrid account of predictability effects in spoken language.

5.5 Future directions

5.5.1 Speech production

This dissertation has only examined the acoustic dimensions of word duration, vowel duration, and vowel dispersion (excepting the brief foray into VOT in Chapter 2), but many other aspects of the acoustic signal are relevant for speech production and perception. Future work could examine what other dimensions of variation are implicated in reduction, and to what extent they are relevant for perception.
The role of articulation in reduction is also important and understudied. Studies of speech articulation have revealed that phenomena with similar acoustic consequences can arise from quite different articulatory processes, suggesting differing underlying mechanisms (Edwards et al., 1991). This finding suggests that purely acoustic studies may miss important generalizations. Additionally, any theory which makes reference to notions of articulatory ease or effort must necessarily be compatible with articulatory facts.

A related question is consistency within and between individuals. Does a talker consistently reduce in the same way in response to the same phenomenon? Do some talkers, say, reduce spectrally while others reduce temporally? These kinds of questions can challenge models of speech production where certain acoustic effects are ‘hard-wired’ into the theory, but may also lead to exciting discoveries.

5.5.2 Individual differences

This dissertation examined theory of mind (ToM) as the main dimension of individual difference between individuals, and several effects were observed. In perception, listeners with poor ToM were observed to attend to acoustic cues rather than contextual or lexical cues relatively more than listeners with better ToM, a result consistent with findings of enhanced perception of fine detail in the broader autism phenotype (Mottron et al., 2006). In production, talkers with poor ToM exhibited a greater degree of reduction due to semantic predictability than talkers with better ToM. This latter result cannot clearly be linked to ToM, and it is possible that some other correlated dimension of cognitive or personality variation is responsible. More generally, the role of this kind of individual variability in speech production and perception remains understudied. Measurable personality traits that have at least some theoretical
grounding in being potentially relevant for speech production and perception include interpersonal orientation (Swap and Rubin, 1983) and self-monitoring (Snyder, 1974), which are measures of how one’s behavior relates to others and oneself, respectively. Street and Murphy (1987) offered preliminary evidence that interpersonal orientation influences conversation behavior, such as speech rate, interruption frequency, and turn duration, suggesting that further investigation of these variables of personal variation is warranted. The self-monitoring scale in particular would appear to have an intuitive link to the phenomenon of sociolinguistic accommodation, although I know of no studies to date which explore this question. A similar question, however, was examined by Aguilar et al. (2015), who investigated the role of rejection sensitivity (Downey and Feldman, 1996)—a measure of the extent to which an individual anticipates, perceives, and negatively reacts to social rejection—in mediating speech accommodation in dyadic interactions. Their results suggested that individuals who are highly sensitive to rejection tend to accommodate more to their conversational partner than less sensitive individuals, although the degree to which the interlocutors’ rejection sensitivity were similar or dissimilar also influenced the degree of accommodation. Taken together, Aguilar et al. (2015) and Street and Murphy (1987) offer compelling evidence that personality factors can have a powerful influence on behaviors often considered purely (socio)linguistic. Further questions on this topic for future research include the extent to which these speech behaviors are static functions of personality, and whether they can be influenced by interventions (see e.g. Aronson et al., 2002; Edlund, 1972).
5.5.3 Quantification of predictability

An issue that has come up in all chapters of this dissertation is how exactly predictability should be quantified. As demonstrated in Chapter 4, lexical frequency and neighborhood density are not independent of each other, and there are assumptions behind the calculation of neighborhood density which are acknowledged to be false—namely, that all phonemes are equally similar. The proposal outlined in Chapter 4, of a psychoacoustically-motivated measure of neighborhood, which treats similarity as a continuous rather than binary measure, will go some way to resolving this problem.

Similarly, the concept of ‘discourse mention’ is more complex than simply counting the number of times a wordform has been uttered in a conversation (Bard and Anderson, 1994; Bard et al., 1989; Beaver et al., 2007), although it is not entirely clear how pragmatic information can be mapped to a quantification of predictability. These issues are left for future research.

5.5.4 Communication

Finally, the question of communication is of central importance to the listener-oriented and passive evolutionary accounts of phonetic reduction. The role of communication in the functional organization of language is still not fully understood. Likewise, the extent to which communication drives language change and historical processes is not known. Language has non-communicative functions in addition to communicative functions, and it is reasonable to ask whether these other functions also influence the use and development of language. Answers to these questions will necessarily have to come from a variety of domains, including sociolinguistics, pragmatics, historical linguistics, and human evolution. Such an enterprise clearly calls for interdisciplinary dialogue to solve the problem of reduction.
Bibliography


292


Vajrabhaya, P. and Kapatsinski, V. (2011). There is more to the story: First-mention
lengthening in Thai interactive discourse. In Proceedings of the Seventeenth Interna-
tional Congress of the Phonetic Sciences.

Speech Communication, 28(2):125–140.


tervocalic consonants: A case for efficient communication. Speech Communication,
47:100–123.

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., and
Preti, A. (2013). The “reading the mind in the eyes” test: Systematic review of
psychometric properties and a validation study in Italy. Cognitive Neuropsychiatry,

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on
speech production. Journal of Experimental Psychology: Learning, Memory and Cog-
nition, 28:735–747.

activation in spoken word recognition. Journal of Memory and Language, 40:374–
408.

Vitevitch, M. S. and Rodríguez, E. (2004). Neighborhood density effects in spoken word
recognition in Spanish. Journal of Multilingual Communication Disorders, 3(1):64–
73.


(and neighborhood frequency) in Spanish speech production: A follow-up report.
University of Kansas Spoken Language Laboratory Technical Report, 1:1–6.

tition: Longer is the same as shorter. Language and Speech, 51:359–381.


Appendix A

_Paired HP and LP SPIN sentences from experiments 1a and 1b_

The nurse gave him first aid
Mr Smith spoke about the aid
A chimpanzee is an ape
She might have discussed the ape
The boat sailed across the bay
Mr Smith knew about the bay
Ruth had a necklace of glass beads
Tom has been discussing the beads
The chicken pecked corn with its beak
She’s glad Bill called about the beak
Bob was cut by the jacknife’s blade
Mary hasn’t discussed the blade
I made the phone call from a booth
Tom heard Jane called about the booth
Tom fell down and got a bad bruise
Sue was interested in the bruise
For your birthday I baked a cake
Tom wants to know about the cake
Greet the heroes with loud cheers

We are considering the cheers
We heard the ticking of the clock
Tom is considering the clock
The detectives searched for a clue
The man spoke about the clue
Harry slept on the folding cot
Tom will discuss the cot
The fruit was shipped in wooden crates
We’ve been discussing the crates
The ship’s Captain summoned his crew
She wants to talk about the crew
The farmer harvested his crop
I want to know about the crop
The steamship left on a cruise
Mr White discussed the cruise
The wedding banquet was a feast
We could consider the feast
The doctor charged a low fee
Tom is talking about the fee
The Admiral commands the fleet
Mr Black considered the fleet
The shepherds guarded their flock
Paul should have discussed the flock
We saw a flock of wild geese
You'd been considering the geese
It was stuck together with glue
Tom has not considered the glue
Lubricate the car with grease
You cannot have discussed the grease
The farmer baled the hay
Tom discussed the hay
At breakfast he drank some juice
We should have considered the juice
This key won't fit in the lock
We hear you called about the lock
The burglar escaped with the loot
Paul hopes we heard about the loot
The lonely bird searched for its mate
I hope Paul asked about the mate
The plow was pulled by an ox
The man should discuss the ox
The story had a clever plot

You're discussing the plot
That accident gave me a scare
Miss Smith considered the scare
Watermelons have lots of seeds
You've considered the seeds
The shepherd watched his flock of sheep
They've considered the sheep
She made the bed with clean sheets
We're discussing the sheets
His boss made him work like a slave
You're glad they heard about the slave
The sport shirt has short sleeves
Nancy has considered the sleeves
Kill the bugs with this spray
Mary had considered the spray
Ruth poured herself a cup of tea
Miss White things about the tea
The house was robbed by a thief
The old woman discussed the thief
The bread was made from whole wheat
We can't consider the wheat
Appendix B

*Paragraph stimulus materials from experiments 1a and 1b*

Target words are indicated in *bold italics*.

Paragraph 1:

In today’s show we will learn how to make a *soup* with *beets* and *green beans*. The first step is to wash and slice the *beets* and *green beans*. Next you fry the onions and toss a sprig of fresh *basil* into the *pot*. Make sure the sprig of *basil* isn’t too big, or its strong *flavor* could overpower the subtle *flavor* from the green beans. Finally, add the beets and string beans, and pour the *meat stock* through a sieve into the *pot*. Nearly any type of *meat stock* will do, but an old chicken carcass boiled down to meat stock works best. Let it simmer for an hour, and at the end you will have a fantastic *soup* that your whole family will enjoy!

Paragraph 2:

My cousin *Sue* competes in winter sports, and was eager to win first place at the *ski meet* this year. She had been training hard after her broken leg healed. Her nemesis *Bobbie* had caused her broken leg just before the *ski meet* last year in a suspicious collision, although *Sue* couldn’t prove it wasn’t an accident. This year, Sue was taking no chances. When *Bobbie* *skied* near enough, Sue threw a pole out in front of her. Bobbie *skied*
straight into a tree in confusion and broke her arm. Best of all, when they investigated the regrettable collision, everyone thought it was an accident!

Paragraph 3:

During his sophomore year of college, Bob Andrews was having a really hard time. He was flunking all his classes, and Top Ramen noodles were all that he could afford. He also broke up with his girlfriend. Then one day, he decided to become a botany major. Suddenly everything changed. He loved the botany classes, and started getting all As. He worked in a research center for a semester, where he earned enough money to swear off noodles and instead buy steak every week. A girl named Judy worked in the research center too, and he was instantly attracted to her. He got close to her by inviting her over for steak at his house every week. After only a couple of these dates she became his girlfriend. Last year, Bob and Judy got married. Bob's life is great, and he owes it all to botany.

Paragraph 4:

On her birthday, Dottie showed up early at the zoo for work. She was in charge of the birds and the baby animals. The emus and the geese were her favorite birds, and the zebras and the otters were her favorite baby animals. She went to the bird house first, and was surprised to see that the emus and the geese were missing and her assistant Lucy wasn’t there. She hurried over to the baby animal area and was upset to find that the zebras and the otters were gone too, and her other assistant Lisa was nowhere to be found. She muttered, “I can’t run the zoo by myself”, and ran back to her office to find out what was going on. But when she opened the door, Dottie heard a honk and a bunch of voices yelling “Surprise!” She realized with joy that not only had Lucy conspired with Lisa to throw a surprise
party for her birthday, but they had even brought along all of her favorite animals!

Paragraph 5:

If you want to go to *Gina’s Pizza Shop*, I can tell you the best way to get there. Go *straight* down this *street* and follow the signs for the *Johnson* Expressway. However, don’t actually go onto the *Johnson* Expressway. When you get to the on-ramp, take a left onto the main *street* in town, *Cleveland* Street. You’ll go past a big school called *Cleveland* High School, right between a church with a yellow door and a church with a *blue steeple*. There is a small alley just past the church with the *blue steeple*. Take this alley for several *blocks*, and turn left on the third road you come to. Eventually, the road will split in two. Take Fillmore Boulevard, which is the one *straight* ahead. About two and a half *blocks* later you’ll see the sign for *Gina’s Pizza Shop*, also known as the best pizza place in town.
Appendix C

Revised Strange Stories task materials

C.1 Mental State Stories

Matthew is a big liar. Matthew’s brother Justin knows this, he knows that Matthew never tells the truth! Yesterday Matthew stole Justin’s ping pong paddle, and Justin knows Matthew has hidden it somewhere, though he can’t find it. He’s very angry. So he finds Matthew and he says, “Where is my ping pong paddle? You must have hidden it either in the closet or under your bed, because I’ve looked everywhere else. Where is it, in the closet or under your bed”? Matthew tells him the paddle is under his bed.

Question: Why will Justin look in the closet for the paddle?

During the war, the Red army captures a member of the Blue army. They want him to tell them where his army’s tanks are; they know they are either by the sea or in the mountains. They know that the prisoner will not want to tell them, he will want to save his army, and so he will certainly lie to them. The prisoner is very brave and very clever, he will not let them find his tanks. The tanks are really in the mountains. When the other side asks him where his tanks are, he says, “They are in the mountains.”

Question: Why did the prisoner say that?
Kyle is always hungry. Today at school it is his favorite meal—fish tacos. He is very greedy, and he would like to have more tacos than anybody else, even though his mother will have a delicious meal for him when he gets home! But everyone is allowed two tacos and no more. When it is Kyle’s turn to be served, he says, “Oh, please can I have four tacos, because I won’t get to eat when I get home!”

Question: Why does Kyle say this?

Samantha wanted to buy a kitten, so she went to see Mrs. Smith, who had lots of kittens she didn’t want. Mrs. Smith loved the kittens, and she wouldn’t do anything to harm them, though she couldn’t keep them all herself. When Samantha visited she wasn’t sure she wanted one of Mrs. Smith’s kittens, since they were all males and she wanted a female. But Mrs. Smith said, “If no one buys the kittens I’ll just have to drown them!”

Question: Why did Mrs. Smith say that?

One day Aunt Michelle came to visit Ryan. Ryan loves his aunt very much, but today she is wearing a new hat; a new hat which Ryan thinks is super ugly. Ryan thinks his aunt looks silly in it, and looks much better in her old hat. But when Aunt Michelle asks Ryan, “How do you like my new hat?” Ryan says, “Oh, it’s very nice.”

Question: Why does he say that?
Rachel waited all year for Christmas, because she knew at Christmas she could ask her parents for a rabbit. Rachel wanted a rabbit more than anything in the world. At last Christmas Day arrived, and Rachel ran to unwrap the big box her parents had given her. She felt sure it would contain a little rabbit in a cage. But when she opened it, with all the family standing around, she found her present was just a boring old set of encyclopedias, which Rachel did not want at all! Still, when Rachel's parents asked her how she liked her Christmas present, she said, “It’s great, thank you. It’s just what I wanted.”

Question: Why did she say this?

Late one night old Mrs. Peabody is walking home. She doesn’t like walking home alone in the dark because she is always afraid that someone will attack her and rob her. She is a very nervous person! Suddenly, a man comes out of the shadows. He wants to ask Mrs. Peabody what time it is, so he walks toward her. When Mrs. Peabody sees the man coming toward her, she starts to tremble and says, “Take my purse, just don’t hurt me please!”

Question: Why did she say that?

A burglar who has just robbed a store is making his getaway. As he is running home, a policeman sees him drop his glove. He doesn’t know the man is a burglar, he just wants to tell him he dropped his glove. But when the policeman shouts out to the burglar, “Hey, you! Stop!” the burglar turns round, sees the policeman and gives himself up. He puts his hands up and admits that he robbed the local store.

Question: Why did the burglar do that?
C.2 Human physical state stories

Bob and Jim are best friends. They are both 10 years old. Bob has brown hair, green eyes and is over 5 feet tall. Jim looks very different from Bob. He has blonde hair and blue eyes and he is much smaller than Bob. Bob and Jim go to the state fair. They go on lots of rides. For the last ride of the day they decide to go on the big rollercoaster. But there is a sign which says: For safety reasons no persons under 5 feet are allowed on.

Question: Why does only Bob go on the rollercoaster?

Robert has never been skiing before and is looking forward to his first skiing vacation this winter. All his gear for the vacation has been well prepared; his mom has bought him a pair of goggles and she has thoroughly waxed and polished the bottom of his skis to protect them. On the first day of Robert’s holiday his skis keep slipping from underneath him, making him fall over into the snow.

Question: Why does Robert keep falling over?

Claire is having her room redecorated; her mother is painting the walls and having new curtains hung. Before, Claire’s room was pink and white with thin net curtains but now the walls are dark red, and brand new thick and expensive velvet curtains have been put up. On the first morning in her new room, Claire fails to wake up at the normal time. As her mother rushes to get her out of bed for school, Claire says it must be too early to get up because it “feels like the middle of the night.”

Question: Why did Claire oversleep?
Sam decides to go on a long walk to get some fresh air. Unfortunately, just after leaving the house, the wind begins to pick up and it starts to rain. Luckily Sam always has an umbrella with him. He quickly opens up the umbrella and wraps his coat tightly around him. Suddenly a gust of wind blows the umbrella straight out of Sam’s hand and it lands in a large bush with lots of thorns. Sam manages to run and grab it before it blows away again and is pleased to find it all in one piece. As he walks home, he notices that his head is starting to get wet despite the umbrella.

Question: Why is Sam getting wet?
### C.3 Physical state stories

Two enemy powers have been at war for a very long time. Each army has won several battles, and the outcome could go either way. The forces are equally matched. However, the Blue army is stronger than the Yellow army in foot soldiers and artillery. But the Yellow army is stronger than the Blue Army in air power. On the day of the final battle, which will decide the outcome of the war, there is heavy fog over the mountains where the fighting is about to occur. Low-lying clouds hang above the soldiers. By the end of the day the Blue army has won.

**Question:** Why did the Blue army win?

A burglar is about to break into a jewelry store. He skillfully picks the lock on the door. Carefully he steps over the electronic detector beam. If he breaks this beam it will set off the alarm. Quietly he opens the door of the storeroom and sees the gems glittering. As he reaches out, however, he steps on something soft. He hears a screech and something small and furry runs out past him, toward the front door. Immediately the alarm sounds.

**Question:** Why did the alarm go off?

Mrs. Robinson is very old and frail. One day she slips on her icy door step and falls on her side. She gets up right away, although she feels quite bruised and shaken. The next day her leg feels very stiff and she can barely walk. She goes to the doctor. As soon as the doctor hears about the fall, and sees her swollen side, he says, “Go to the hospital, right now.” At the hospital they take an X-ray.

**Question:** Why did they take an X-ray?
John is going shopping. He buys a nice new desk lamp for his office. He needs a light bulb for his new lamp. He goes from the furniture department to the electrical department. In the electrical department he finds that there are two brands of the right kind of light bulb. Everbrite light bulbs cost less in single packs than Literite bulbs. However, only Literite bulbs come in multipacks of six. John buys the multipack, even though he only needs one bulb.

Question: Why does John buy the Literite bulbs?

Mrs. Simpson, the librarian, receives a special book which she has to catalogue and find an appropriate place for. She has to decide which section to file it under. The library is very big, and has different sections on many different subjects. The new book is about plants and their medical uses, and is heavily illustrated. However, Mrs. Simpson does not put it on the shelf with the rest of the books on botany. Neither does she put it with the books on medicine. Instead, she carefully takes it into a separate room. In this room all the books are kept in special cases, and the temperature is kept constant.

Question: Why did she do this?

Henry is preparing for a big dinner party. He is famous for his excellent mayonnaise. He has bought lots of fresh eggs. The recipe says, “Carefully separate the yolks of six eggs and add oil very gradually.” He has already bought enough dessert to feed everyone easily. However, he now looks up the recipe for meringue. Henry will not waste anything.

Question: Why does Henry make meringue?
Ethan is very rich, and today he is going to buy an expensive new car. He is considering whether to make a single payment, or whether to spread the cost over the year. If he pays in monthly installments, the dealer will charge 5% interest on the loan. His bank currently gives him 8% interest on the money in his account. Even though he has easily enough money to pay the full amount, he decides to pay by monthly installments.

Question: Why does he do that?

Sara is very far-sighted. She has only one pair of glasses, which she keeps losing. Today she has lost her glasses again and she needs to find them. She had them yesterday evening when she looked up the television programs. She must have left them somewhere that she has been today. She asks Nick to find her glasses. She tells him that today she went to her regular early morning exercise class, then to the post office, and last to the florist. Nick goes straight to the post office.

Question: Why is the post office the most likely place to look?
### C.4 Unlinked stories

<table>
<thead>
<tr>
<th>The two countries had been at war. A housewife is about to enter the super-market. Today he is going to buy an expensive new stereo. Mrs. Brown, the postal worker, receives a special package. Mrs. Pearson wouldn't hurt a fly. Mary's birthday is in February. Late one evening the old man was watching television.</th>
<th>Question: When is Mary's birthday?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young Simon is very robust. She sees that Fred cannot play. Jeremy is always laughing. Ruth sees her uncle very often, but today he has gone to Brazil. Richard is packing up to go away. Today, at college, it is Jim's worst class—statistical mechanics. She has only one dollar left, which she must keep for her bus fare. He buys a bright tie, to go with his new shirt.</td>
<td>Question: How many dollars does she have left?</td>
</tr>
<tr>
<td>Simon takes the special butter from the refrigerator. Each boxer has won several fights. He skillfully picks out the imperfect items. They are either in Boston or in New York. She has to cut the grass and find somewhere to plant the bay tree. The conductor sees that the cellist has broken a string. Heather took the bus to the station.</td>
<td>Question: Where did Heather take the bus to?</td>
</tr>
<tr>
<td>The four brothers stood aside to make room for their sister, Amy. Scott repeated the experiment, several times. The name of the airport has changed. Louise uncorked a little bottle of oil. The 2 children had to abandon their daily walk. She took a suite in a grand hotel. It had already been 20 years since the operation.</td>
<td>Question: Who abandoned their daily walk?</td>
</tr>
</tbody>
</table>
One day Uncle Simon came to visit Alex. The first part of the performance had come to an end. He put away the letter and stuck his hands in his pockets. She was still holding her umbrella. The cats ran back to the boy. Flora came into the middle of the square. The little island had a high rocky shoreline.

Question: Where did Flora go to?

At the edge of the road a little grass was growing. He reaches out to find the light switch. A sailor who has just left his ship is walking to the town. She has to decide where to keep the pasta. At last daylight came, and Tommy got out of bed to open his presents. Jim knows all about investing money, as he works in a large bank. They exchanged a few brief words about the weather.

Question: Why did Tommy get out of bed?

She is always saying that someone will eventually find the treasure. Everyone is allowed two visits and no more. At the psychiatry department they were interviewing the new nurses. Jim will win the first race of the meet. She has taken all the children to visit the zoo today. Simon’s uncle is wearing a new suit. The same phrase of twenty-three notes recurred throughout.

Question: What will Jim win?

He needs a new engine for his old car. The prize is an immediate lump sum of $20,000 tax-free. Japan is economically stronger than Italy. The mother is very brave and long suffering. The new book is about statistics and experimental design, and contains many graphs. The front room contained a little bird in a cage. Although Jim is only 29 years old, he has an income of $150,000 per year. There are not many people in the large rectangular dining room this evening.

Question: Who is brave and long suffering?
Appendix D

*Stimulus materials used in experiment 2*

Both lists saw the same targets and fillers in the same orders. The lists differed in the distractor items, which were either genuine fillers or minimal pair neighbors.

<table>
<thead>
<tr>
<th>Target word</th>
<th>List 1 distractor</th>
<th>List 2 distractor</th>
<th>Filler</th>
</tr>
</thead>
<tbody>
<tr>
<td>punk</td>
<td>bunk</td>
<td>date</td>
<td>league</td>
</tr>
<tr>
<td>pork</td>
<td>soil</td>
<td>soil</td>
<td>brick</td>
</tr>
<tr>
<td>posh</td>
<td>witch</td>
<td>witch</td>
<td>mood</td>
</tr>
<tr>
<td>torch</td>
<td>fence</td>
<td>fence</td>
<td>stress</td>
</tr>
<tr>
<td>pipe</td>
<td>text</td>
<td>text</td>
<td>tide</td>
</tr>
<tr>
<td>palm</td>
<td>noise</td>
<td>balm</td>
<td>square</td>
</tr>
<tr>
<td>pinch</td>
<td>bell</td>
<td>bell</td>
<td>inch</td>
</tr>
<tr>
<td>tooth</td>
<td>flesh</td>
<td>flesh</td>
<td>wealth</td>
</tr>
<tr>
<td>pig</td>
<td>big</td>
<td>bait</td>
<td>spot</td>
</tr>
<tr>
<td>cob</td>
<td>date</td>
<td>gob</td>
<td>fund</td>
</tr>
<tr>
<td>kilt</td>
<td>hole</td>
<td>guilt</td>
<td>bomb</td>
</tr>
<tr>
<td>test</td>
<td>ward</td>
<td>ward</td>
<td>thigh</td>
</tr>
<tr>
<td>tile</td>
<td>dial</td>
<td>rum</td>
<td>cloth</td>
</tr>
<tr>
<td>pan</td>
<td>sand</td>
<td>ban</td>
<td>ice</td>
</tr>
<tr>
<td>peach</td>
<td>shirt</td>
<td>beach</td>
<td>cloud</td>
</tr>
<tr>
<td>Target word</td>
<td>List 1 distractor</td>
<td>List 2 distractor</td>
<td>Filler</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------</td>
<td>------------------</td>
<td>--------</td>
</tr>
<tr>
<td>pup</td>
<td>van</td>
<td>van</td>
<td>gate</td>
</tr>
<tr>
<td>teal</td>
<td>deal</td>
<td>hair</td>
<td>hook</td>
</tr>
<tr>
<td>tights</td>
<td>church</td>
<td>church</td>
<td>sale</td>
</tr>
<tr>
<td>peal</td>
<td>dance</td>
<td>dance</td>
<td>stick</td>
</tr>
<tr>
<td>tempt</td>
<td>bone</td>
<td>bone</td>
<td>chief</td>
</tr>
<tr>
<td>coot</td>
<td>drive</td>
<td>drive</td>
<td>cave</td>
</tr>
<tr>
<td>pomp</td>
<td>moon</td>
<td>moon</td>
<td>smoke</td>
</tr>
<tr>
<td>paunch</td>
<td>heel</td>
<td>heel</td>
<td>golf</td>
</tr>
<tr>
<td>cake</td>
<td>bush</td>
<td>bush</td>
<td>waste</td>
</tr>
<tr>
<td>torque</td>
<td>dork</td>
<td>ship</td>
<td>rod</td>
</tr>
<tr>
<td>toe</td>
<td>doe</td>
<td>noise</td>
<td>lake</td>
</tr>
<tr>
<td>poach</td>
<td>fight</td>
<td>fight</td>
<td>beer</td>
</tr>
<tr>
<td>kin</td>
<td>fuel</td>
<td>fuel</td>
<td>toy</td>
</tr>
<tr>
<td>tap</td>
<td>stream</td>
<td>stream</td>
<td>wheat</td>
</tr>
<tr>
<td>torn</td>
<td>wheel</td>
<td>wheel</td>
<td>fee</td>
</tr>
<tr>
<td>pad</td>
<td>land</td>
<td>bad</td>
<td>fool</td>
</tr>
<tr>
<td>kiln</td>
<td>bowl</td>
<td>bowl</td>
<td>queen</td>
</tr>
<tr>
<td>coo</td>
<td>guest</td>
<td>goo</td>
<td>nurse</td>
</tr>
<tr>
<td>cud</td>
<td>screen</td>
<td>screen</td>
<td>pair</td>
</tr>
<tr>
<td>tuck</td>
<td>duck</td>
<td>mark</td>
<td>joke</td>
</tr>
<tr>
<td>part</td>
<td>grain</td>
<td>grain</td>
<td>steps</td>
</tr>
<tr>
<td>pith</td>
<td>map</td>
<td>map</td>
<td>vote</td>
</tr>
<tr>
<td>cape</td>
<td>bait</td>
<td>gape</td>
<td>chest</td>
</tr>
<tr>
<td>cuff</td>
<td>hair</td>
<td>guff</td>
<td>band</td>
</tr>
<tr>
<td>Target word</td>
<td>List 1 distractor</td>
<td>List 2 distractor</td>
<td>Filler</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------</td>
<td>------------------</td>
<td>--------</td>
</tr>
<tr>
<td>tart</td>
<td>dart</td>
<td>hand</td>
<td>cheek</td>
</tr>
<tr>
<td>pal</td>
<td>boot</td>
<td>boot</td>
<td>snow</td>
</tr>
<tr>
<td>kit</td>
<td>hair</td>
<td>git</td>
<td>prince</td>
</tr>
<tr>
<td>pill</td>
<td>bill</td>
<td>book</td>
<td>tool</td>
</tr>
<tr>
<td>car</td>
<td>root</td>
<td>root</td>
<td>tent</td>
</tr>
<tr>
<td>cop</td>
<td>cash</td>
<td>cash</td>
<td>cream</td>
</tr>
<tr>
<td>cog</td>
<td>sheet</td>
<td>sheet</td>
<td>stock</td>
</tr>
<tr>
<td>pox</td>
<td>box</td>
<td>salt</td>
<td>cap</td>
</tr>
<tr>
<td>tat</td>
<td>cheese</td>
<td>cheese</td>
<td>seed</td>
</tr>
<tr>
<td>cad</td>
<td>pig</td>
<td>pig</td>
<td>tank</td>
</tr>
<tr>
<td>peace</td>
<td>show</td>
<td>show</td>
<td>clock</td>
</tr>
<tr>
<td>card</td>
<td>book</td>
<td>guard</td>
<td>taste</td>
</tr>
<tr>
<td>coat</td>
<td>dance</td>
<td>goat</td>
<td>tin</td>
</tr>
<tr>
<td>taunt</td>
<td>daunt</td>
<td>hole</td>
<td>tongue</td>
</tr>
<tr>
<td>taint</td>
<td>guy</td>
<td>guy</td>
<td>mark</td>
</tr>
<tr>
<td>pulp</td>
<td>guide</td>
<td>guide</td>
<td>soul</td>
</tr>
<tr>
<td>park</td>
<td>ship</td>
<td>bark</td>
<td>roof</td>
</tr>
<tr>
<td>pep</td>
<td>trout</td>
<td>trout</td>
<td>goal</td>
</tr>
<tr>
<td>cab</td>
<td>bag</td>
<td>gab</td>
<td>bath</td>
</tr>
<tr>
<td>tab</td>
<td>dab</td>
<td>fish</td>
<td>breathing</td>
</tr>
<tr>
<td>comb</td>
<td>aid</td>
<td>aid</td>
<td>twin</td>
</tr>
<tr>
<td>core</td>
<td>gun</td>
<td>gore</td>
<td>tie</td>
</tr>
<tr>
<td>tan</td>
<td>Dan</td>
<td>gun</td>
<td>cell</td>
</tr>
<tr>
<td>teat</td>
<td>deer</td>
<td>deer</td>
<td>file</td>
</tr>
<tr>
<td>Target word</td>
<td>List 1 distractor</td>
<td>List 2 distractor</td>
<td>Filler</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------</td>
<td>------------------</td>
<td>--------</td>
</tr>
<tr>
<td>punch</td>
<td>bunch</td>
<td>dance</td>
<td>pool</td>
</tr>
<tr>
<td>teethe</td>
<td>phone</td>
<td>phone</td>
<td>goods</td>
</tr>
<tr>
<td>putt</td>
<td>butt</td>
<td>dice</td>
<td>cut</td>
</tr>
<tr>
<td>pub</td>
<td>cat</td>
<td>cat</td>
<td>rope</td>
</tr>
<tr>
<td>tote</td>
<td>dote</td>
<td>shirt</td>
<td>song</td>
</tr>
<tr>
<td>cod</td>
<td>dice</td>
<td>god</td>
<td>pot</td>
</tr>
<tr>
<td>curb</td>
<td>priest</td>
<td>priest</td>
<td>truck</td>
</tr>
<tr>
<td>pants</td>
<td>frame</td>
<td>frame</td>
<td>throat</td>
</tr>
<tr>
<td>tuft</td>
<td>plane</td>
<td>plane</td>
<td>jail</td>
</tr>
<tr>
<td>cast</td>
<td>trick</td>
<td>trick</td>
<td>cake</td>
</tr>
<tr>
<td>curl</td>
<td>job</td>
<td>girl</td>
<td>tail</td>
</tr>
<tr>
<td>pall</td>
<td>rum</td>
<td>ball</td>
<td>faith</td>
</tr>
<tr>
<td>type</td>
<td>steel</td>
<td>steel</td>
<td>pile</td>
</tr>
<tr>
<td>tank</td>
<td>dank</td>
<td>job</td>
<td>shell</td>
</tr>
<tr>
<td>peat</td>
<td>beat</td>
<td>bag</td>
<td>flight</td>
</tr>
<tr>
<td>toast</td>
<td>gift</td>
<td>gift</td>
<td>fly</td>
</tr>
<tr>
<td>tiff</td>
<td>crowd</td>
<td>crowd</td>
<td>crime</td>
</tr>
<tr>
<td>cub</td>
<td>knife</td>
<td>knife</td>
<td>chain</td>
</tr>
<tr>
<td>tame</td>
<td>dame</td>
<td>guest</td>
<td>port</td>
</tr>
<tr>
<td>code</td>
<td>fish</td>
<td>goad</td>
<td>wing</td>
</tr>
<tr>
<td>cone</td>
<td>fate</td>
<td>fate</td>
<td>match</td>
</tr>
<tr>
<td>corn</td>
<td>plate</td>
<td>plate</td>
<td>path</td>
</tr>
<tr>
<td>pulse</td>
<td>sheep</td>
<td>sheep</td>
<td>ring</td>
</tr>
<tr>
<td>tomb</td>
<td>doom</td>
<td>sand</td>
<td>ash</td>
</tr>
<tr>
<td>Target word</td>
<td>List 1 distractor</td>
<td>List 2 distractor</td>
<td>Filler</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------</td>
<td>------------------</td>
<td>----------</td>
</tr>
<tr>
<td>tag</td>
<td>claim</td>
<td>claim</td>
<td>storm</td>
</tr>
<tr>
<td>peak</td>
<td>mark</td>
<td>beak</td>
<td>theme</td>
</tr>
<tr>
<td>toad</td>
<td>block</td>
<td>block</td>
<td>tape</td>
</tr>
<tr>
<td>coast</td>
<td>salt</td>
<td>ghost</td>
<td>stairs</td>
</tr>
<tr>
<td>tense</td>
<td>dense</td>
<td>land</td>
<td>fruit</td>
</tr>
</tbody>
</table>
Appendix E

Word stimuli used in experiment 3

<table>
<thead>
<tr>
<th>Word</th>
<th>Transcription</th>
<th>Log frequency</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>act</td>
<td>/ækt/</td>
<td>3.4518</td>
<td>9</td>
</tr>
<tr>
<td>ask</td>
<td>/æsk/</td>
<td>3.1072</td>
<td>8</td>
</tr>
<tr>
<td>axe</td>
<td>/æks/</td>
<td>1.7782</td>
<td>9</td>
</tr>
<tr>
<td>back</td>
<td>/bæk/</td>
<td>3.9854</td>
<td>32</td>
</tr>
<tr>
<td>bad</td>
<td>/bæd/</td>
<td>3.1553</td>
<td>30</td>
</tr>
<tr>
<td>bag</td>
<td>/bæg/</td>
<td>2.6232</td>
<td>28</td>
</tr>
<tr>
<td>bath</td>
<td>/bæθ/</td>
<td>2.415</td>
<td>17</td>
</tr>
<tr>
<td>black</td>
<td>/blaek/</td>
<td>3.3075</td>
<td>10</td>
</tr>
<tr>
<td>brag</td>
<td>/braeg/</td>
<td>1.301</td>
<td>11</td>
</tr>
<tr>
<td>cat</td>
<td>/kæt/</td>
<td>2.3617</td>
<td>35</td>
</tr>
<tr>
<td>catch</td>
<td>/kætʃ/</td>
<td>2.6335</td>
<td>20</td>
</tr>
<tr>
<td>class</td>
<td>/klæs/</td>
<td>3.316</td>
<td>12</td>
</tr>
<tr>
<td>crab</td>
<td>/kræb/</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>crash</td>
<td>/kræʃ/</td>
<td>2.301</td>
<td>13</td>
</tr>
<tr>
<td>draft</td>
<td>/draeft/</td>
<td>2.3802</td>
<td>7</td>
</tr>
<tr>
<td>fact</td>
<td>/fækt/</td>
<td>3.6503</td>
<td>6</td>
</tr>
<tr>
<td>flax</td>
<td>/flæks/</td>
<td>1.4771</td>
<td>5</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>glass</td>
<td>/glæz/</td>
<td>2.9956</td>
<td>7</td>
</tr>
<tr>
<td>gnat</td>
<td>/næt/</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>hat</td>
<td>/hæt/</td>
<td>2.7482</td>
<td>34</td>
</tr>
<tr>
<td>jazz</td>
<td>/dʒæz/</td>
<td>2.9956</td>
<td>7</td>
</tr>
<tr>
<td>mask</td>
<td>/mæsk/</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>match</td>
<td>/mæʧ/</td>
<td>2.6128</td>
<td>14</td>
</tr>
<tr>
<td>math</td>
<td>/mæθ/</td>
<td>1.6021</td>
<td>15</td>
</tr>
<tr>
<td>nap</td>
<td>/næp/</td>
<td>1.6021</td>
<td>20</td>
</tr>
<tr>
<td>pact</td>
<td>/pækt/</td>
<td>2.3802</td>
<td>10</td>
</tr>
<tr>
<td>pass</td>
<td>/pæs/</td>
<td>2.9494</td>
<td>24</td>
</tr>
<tr>
<td>path</td>
<td>/pæθ/</td>
<td>2.6435</td>
<td>14</td>
</tr>
<tr>
<td>raft</td>
<td>/ræft/</td>
<td>1.6021</td>
<td>14</td>
</tr>
<tr>
<td>rat</td>
<td>/ræt/</td>
<td>1.7782</td>
<td>37</td>
</tr>
<tr>
<td>sad</td>
<td>/sæd/</td>
<td>2.5441</td>
<td>25</td>
</tr>
<tr>
<td>scab</td>
<td>/skæb/</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>scrap</td>
<td>/skræp/</td>
<td>1.9031</td>
<td>4</td>
</tr>
<tr>
<td>scratch</td>
<td>/skræʧ/</td>
<td>1.9542</td>
<td>2</td>
</tr>
<tr>
<td>slap</td>
<td>/slæp/</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>snag</td>
<td>/snæg/</td>
<td>1.4771</td>
<td>9</td>
</tr>
<tr>
<td>splash</td>
<td>/splæʧ/</td>
<td>1.4771</td>
<td>2</td>
</tr>
<tr>
<td>staff</td>
<td>/stæf/</td>
<td>3.0531</td>
<td>5</td>
</tr>
<tr>
<td>tag</td>
<td>/tæg/</td>
<td>1.699</td>
<td>21</td>
</tr>
<tr>
<td>task</td>
<td>/tæsk/</td>
<td>2.7782</td>
<td>9</td>
</tr>
<tr>
<td>tax</td>
<td>/tæks/</td>
<td>3.2945</td>
<td>9</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>---------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>track</td>
<td>/træk/</td>
<td>2.5798</td>
<td>10</td>
</tr>
<tr>
<td>boss</td>
<td>/bəs/</td>
<td>2.301</td>
<td>11</td>
</tr>
<tr>
<td>bought</td>
<td>/bɔt/</td>
<td>2.7482</td>
<td>25</td>
</tr>
<tr>
<td>broth</td>
<td>/brɔθ/</td>
<td>1.4771</td>
<td>8</td>
</tr>
<tr>
<td>caught</td>
<td>/kɔt/</td>
<td>2.9912</td>
<td>23</td>
</tr>
<tr>
<td>cause</td>
<td>/kɔz/</td>
<td>3.1139</td>
<td>10</td>
</tr>
<tr>
<td>chalk</td>
<td>/tʃɔk/</td>
<td>1.4771</td>
<td>13</td>
</tr>
<tr>
<td>clause</td>
<td>/klɔz/</td>
<td>1.9542</td>
<td>4</td>
</tr>
<tr>
<td>claws</td>
<td>/klɔz/</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>cloth</td>
<td>/klɔθ/</td>
<td>2.6335</td>
<td>3</td>
</tr>
<tr>
<td>cost</td>
<td>/kɔst/</td>
<td>3.3598</td>
<td>5</td>
</tr>
<tr>
<td>cough</td>
<td>/kɔf/</td>
<td>1.8451</td>
<td>11</td>
</tr>
<tr>
<td>dog</td>
<td>/dɔg/</td>
<td>2.8751</td>
<td>8</td>
</tr>
<tr>
<td>faun</td>
<td>/fɔn/</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>flaw</td>
<td>/flɔ/</td>
<td>1.4771</td>
<td>9</td>
</tr>
<tr>
<td>floss</td>
<td>/flɔs/</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>fought</td>
<td>/fɔt/</td>
<td>2.6628</td>
<td>20</td>
</tr>
<tr>
<td>frog</td>
<td>/frɔɡ/</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>gauze</td>
<td>/ɡɔz/</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>gone</td>
<td>/ɡɔn/</td>
<td>3.29</td>
<td>17</td>
</tr>
<tr>
<td>haunt</td>
<td>/hɔnt/</td>
<td>1.6021</td>
<td>11</td>
</tr>
<tr>
<td>jaw</td>
<td>/dʒɔ/</td>
<td>2.2041</td>
<td>15</td>
</tr>
<tr>
<td>law</td>
<td>/lɔ/</td>
<td>3.4757</td>
<td>24</td>
</tr>
<tr>
<td>lawn</td>
<td>/lɔn/</td>
<td>2.1761</td>
<td>19</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>long</td>
<td>/lɔŋ/</td>
<td>3.8779</td>
<td>13</td>
</tr>
<tr>
<td>lost</td>
<td>/lɔst/</td>
<td>3.238</td>
<td>9</td>
</tr>
<tr>
<td>moss</td>
<td>/mɔs/</td>
<td>1.9542</td>
<td>15</td>
</tr>
<tr>
<td>moth</td>
<td>/mɔθ/</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>off</td>
<td>/ɔf/</td>
<td>3.8055</td>
<td>12</td>
</tr>
<tr>
<td>sauce</td>
<td>/sɔs/</td>
<td>2.301</td>
<td>10</td>
</tr>
<tr>
<td>sloth</td>
<td>/slɔθ/</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>soft</td>
<td>/soft/</td>
<td>2.7853</td>
<td>5</td>
</tr>
<tr>
<td>smog</td>
<td>/smɔɡ/</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>straw</td>
<td>/strɔ/</td>
<td>2.1761</td>
<td>3</td>
</tr>
<tr>
<td>strong</td>
<td>/strɔŋ/</td>
<td>3.3054</td>
<td>3</td>
</tr>
<tr>
<td>talk</td>
<td>/tɔk/</td>
<td>3.1875</td>
<td>23</td>
</tr>
<tr>
<td>thought</td>
<td>/θɔt/</td>
<td>3.7118</td>
<td>11</td>
</tr>
<tr>
<td>toss</td>
<td>/tɔs/</td>
<td>1.9542</td>
<td>10</td>
</tr>
<tr>
<td>walk</td>
<td>/wɔk/</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>wash</td>
<td>/wɔʃ/</td>
<td>2.5682</td>
<td>7</td>
</tr>
<tr>
<td>watch</td>
<td>/wɔtʃ/</td>
<td>2.9085</td>
<td>5</td>
</tr>
<tr>
<td>water</td>
<td>/wɔtə/</td>
<td>3.6454</td>
<td>3</td>
</tr>
<tr>
<td>bed</td>
<td>/bɛd/</td>
<td>3.1038</td>
<td>25</td>
</tr>
<tr>
<td>best</td>
<td>/bɛst/</td>
<td>3.5453</td>
<td>23</td>
</tr>
<tr>
<td>bet</td>
<td>/bɛt/</td>
<td>2.301</td>
<td>32</td>
</tr>
<tr>
<td>bread</td>
<td>/breð/</td>
<td>2.6232</td>
<td>15</td>
</tr>
<tr>
<td>breast</td>
<td>/brest/</td>
<td>2.0414</td>
<td>6</td>
</tr>
<tr>
<td>breath</td>
<td>/breθ/</td>
<td>2.7243</td>
<td>3</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>chef</td>
<td>/ʃɛf/</td>
<td>1.9542</td>
<td>8</td>
</tr>
<tr>
<td>chess</td>
<td>/tʃɛs/</td>
<td>1.4771</td>
<td>11</td>
</tr>
<tr>
<td>chest</td>
<td>/tʃɛst/</td>
<td>2.7243</td>
<td>13</td>
</tr>
<tr>
<td>deaf</td>
<td>/def/</td>
<td>2.0792</td>
<td>13</td>
</tr>
<tr>
<td>death</td>
<td>/dɛθ/</td>
<td>3.4425</td>
<td>10</td>
</tr>
<tr>
<td>deck</td>
<td>/dek/</td>
<td>2.3617</td>
<td>20</td>
</tr>
<tr>
<td>depth</td>
<td>/dɛpθ/</td>
<td>2.7243</td>
<td>1</td>
</tr>
<tr>
<td>desk</td>
<td>/desk/</td>
<td>2.8129</td>
<td>3</td>
</tr>
<tr>
<td>dress</td>
<td>/dres/</td>
<td>2.8261</td>
<td>9</td>
</tr>
<tr>
<td>flesh</td>
<td>/flɛʃ/</td>
<td>2.716</td>
<td>7</td>
</tr>
<tr>
<td>fresh</td>
<td>/frɛʃ/</td>
<td>2.9138</td>
<td>4</td>
</tr>
<tr>
<td>head</td>
<td>/hɛd/</td>
<td>3.6274</td>
<td>25</td>
</tr>
<tr>
<td>leg</td>
<td>/lɛg/</td>
<td>2.7634</td>
<td>15</td>
</tr>
<tr>
<td>mesh</td>
<td>/mɛʃ/</td>
<td>1.6021</td>
<td>6</td>
</tr>
<tr>
<td>mess</td>
<td>/mes/</td>
<td>2.3424</td>
<td>18</td>
</tr>
<tr>
<td>neck</td>
<td>/nɛk/</td>
<td>2.9085</td>
<td>13</td>
</tr>
<tr>
<td>nest</td>
<td>/nest/</td>
<td>2.301</td>
<td>15</td>
</tr>
<tr>
<td>net</td>
<td>/nɛt/</td>
<td>2.5315</td>
<td>26</td>
</tr>
<tr>
<td>pet</td>
<td>/pet/</td>
<td>1.9031</td>
<td>30</td>
</tr>
<tr>
<td>quest</td>
<td>/kwɛst/</td>
<td>2.2041</td>
<td>2</td>
</tr>
<tr>
<td>rest</td>
<td>/rest/</td>
<td>3.2148</td>
<td>20</td>
</tr>
<tr>
<td>sect</td>
<td>/sɛkt/</td>
<td>1.301</td>
<td>5</td>
</tr>
<tr>
<td>shed</td>
<td>/ʃɛd/</td>
<td>2.0414</td>
<td>17</td>
</tr>
<tr>
<td>sketch</td>
<td>/skɛtʃ/</td>
<td>2.2041</td>
<td>3</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>--------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>sled</td>
<td>/slɛd/</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>step</td>
<td>/stɛp/</td>
<td>3.1173</td>
<td>6</td>
</tr>
<tr>
<td>stress</td>
<td>/strɛs/</td>
<td>3.0294</td>
<td>2</td>
</tr>
<tr>
<td>sweat</td>
<td>/swɛt/</td>
<td>2.3617</td>
<td>9</td>
</tr>
<tr>
<td>test</td>
<td>/tɛst/</td>
<td>3.0755</td>
<td>18</td>
</tr>
<tr>
<td>theft</td>
<td>/θɛft/</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>thread</td>
<td>/θrɛd/</td>
<td>2.1761</td>
<td>6</td>
</tr>
<tr>
<td>threat</td>
<td>/θrɛt/</td>
<td>2.6232</td>
<td>3</td>
</tr>
<tr>
<td>zest</td>
<td>/zɛst/</td>
<td>1.699</td>
<td>11</td>
</tr>
<tr>
<td>beach</td>
<td>/biʃ/</td>
<td>2.8261</td>
<td>18</td>
</tr>
<tr>
<td>beak</td>
<td>/bɪk/</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>beat</td>
<td>/bɪt/</td>
<td>2.8325</td>
<td>33</td>
</tr>
<tr>
<td>beef</td>
<td>/bɪf/</td>
<td>2.5052</td>
<td>15</td>
</tr>
<tr>
<td>bleed</td>
<td>/blɪd/</td>
<td>1.301</td>
<td>9</td>
</tr>
<tr>
<td>brief</td>
<td>/brɪf/</td>
<td>2.8633</td>
<td>8</td>
</tr>
<tr>
<td>cheap</td>
<td>/tʃɪp/</td>
<td>2.3802</td>
<td>18</td>
</tr>
<tr>
<td>cheese</td>
<td>/tʃɪz/</td>
<td>1.9542</td>
<td>13</td>
</tr>
<tr>
<td>chief</td>
<td>/tʃɪf/</td>
<td>3.0755</td>
<td>12</td>
</tr>
<tr>
<td>dream</td>
<td>/drɪm/</td>
<td>2.8062</td>
<td>7</td>
</tr>
<tr>
<td>feast</td>
<td>/fɪst/</td>
<td>1.4771</td>
<td>10</td>
</tr>
<tr>
<td>flea</td>
<td>/fli/</td>
<td>1.4771</td>
<td>22</td>
</tr>
<tr>
<td>fleece</td>
<td>/flɪs/</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>greed</td>
<td>/ɡrɪd/</td>
<td>1.4771</td>
<td>12</td>
</tr>
<tr>
<td>grief</td>
<td>/ɡrɪf/</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>heat</td>
<td>/hit/</td>
<td>2.9868</td>
<td>25</td>
</tr>
<tr>
<td>jeans</td>
<td>/dʒinz/</td>
<td>1.301</td>
<td>4</td>
</tr>
<tr>
<td>knee</td>
<td>/ni/</td>
<td>2.5441</td>
<td>31</td>
</tr>
<tr>
<td>mead</td>
<td>/mid/</td>
<td>1.301</td>
<td>20</td>
</tr>
<tr>
<td>meat</td>
<td>/mit/</td>
<td>3.2856</td>
<td>25</td>
</tr>
<tr>
<td>niece</td>
<td>/nis/</td>
<td>1.9031</td>
<td>14</td>
</tr>
<tr>
<td>please</td>
<td>/pliz/</td>
<td>2.7924</td>
<td>6</td>
</tr>
<tr>
<td>priest</td>
<td>/prist/</td>
<td>2.2041</td>
<td>1</td>
</tr>
<tr>
<td>scheme</td>
<td>/skim/</td>
<td>2.5185</td>
<td>6</td>
</tr>
<tr>
<td>seed</td>
<td>/sid/</td>
<td>2.6128</td>
<td>27</td>
</tr>
<tr>
<td>sheep</td>
<td>/ʃip/</td>
<td>2.3617</td>
<td>20</td>
</tr>
<tr>
<td>sheet</td>
<td>/ʃit/</td>
<td>2.6532</td>
<td>23</td>
</tr>
<tr>
<td>ski</td>
<td>/ski/</td>
<td>1.699</td>
<td>9</td>
</tr>
<tr>
<td>sleeve</td>
<td>/sliv/</td>
<td>2.0414</td>
<td>7</td>
</tr>
<tr>
<td>speech</td>
<td>/spitʃ/</td>
<td>2.7853</td>
<td>3</td>
</tr>
<tr>
<td>speed</td>
<td>/spid/</td>
<td>2.9191</td>
<td>8</td>
</tr>
<tr>
<td>squeak</td>
<td>/skwik/</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>stream</td>
<td>/strim/</td>
<td>2.7076</td>
<td>6</td>
</tr>
<tr>
<td>street</td>
<td>/strit/</td>
<td>3.3874</td>
<td>6</td>
</tr>
<tr>
<td>suite</td>
<td>/swit/</td>
<td>2.9868</td>
<td>9</td>
</tr>
<tr>
<td>teeth</td>
<td>/tiθ/</td>
<td>3.0128</td>
<td>12</td>
</tr>
<tr>
<td>theme</td>
<td>/θim/</td>
<td>2.7404</td>
<td>8</td>
</tr>
<tr>
<td>thief</td>
<td>/θif/</td>
<td>1.9031</td>
<td>8</td>
</tr>
<tr>
<td>treat</td>
<td>/trit/</td>
<td>2.415</td>
<td>10</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>tree</td>
<td>/tri/</td>
<td>2.7709</td>
<td>16</td>
</tr>
<tr>
<td>tweed</td>
<td>/twid/</td>
<td>1.699</td>
<td>4</td>
</tr>
<tr>
<td>tweet</td>
<td>/twit/</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>weed</td>
<td>/wid/</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>wheat</td>
<td>/hwit/</td>
<td>1.9542</td>
<td>9</td>
</tr>
<tr>
<td>yeast</td>
<td>/jist/</td>
<td>1.4771</td>
<td>4</td>
</tr>
<tr>
<td>box</td>
<td>/baks/</td>
<td>2.8451</td>
<td>5</td>
</tr>
<tr>
<td>chop</td>
<td>/tfap/</td>
<td>1.4771</td>
<td>18</td>
</tr>
<tr>
<td>clog</td>
<td>/klog/</td>
<td>1.301</td>
<td>6</td>
</tr>
<tr>
<td>cop</td>
<td>/kap/</td>
<td>2.1761</td>
<td>30</td>
</tr>
<tr>
<td>cot</td>
<td>/kat/</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>crop</td>
<td>/krap/</td>
<td>2.301</td>
<td>9</td>
</tr>
<tr>
<td>dodge</td>
<td>/dadʒ/</td>
<td>2.0414</td>
<td>8</td>
</tr>
<tr>
<td>drop</td>
<td>/drap/</td>
<td>2.7709</td>
<td>6</td>
</tr>
<tr>
<td>fox</td>
<td>/faks/</td>
<td>2.1139</td>
<td>8</td>
</tr>
<tr>
<td>flock</td>
<td>/frak/</td>
<td>1.301</td>
<td>4</td>
</tr>
<tr>
<td>god</td>
<td>/ɡad/</td>
<td>3.5024</td>
<td>19</td>
</tr>
<tr>
<td>hot</td>
<td>/hɒt/</td>
<td>3.1139</td>
<td>28</td>
</tr>
<tr>
<td>job</td>
<td>/dʒəb/</td>
<td>3.3766</td>
<td>19</td>
</tr>
<tr>
<td>knot</td>
<td>/nɒt/</td>
<td>4.6644</td>
<td>27</td>
</tr>
<tr>
<td>lodge</td>
<td>/lɑdʒ/</td>
<td>2.2788</td>
<td>10</td>
</tr>
<tr>
<td>mob</td>
<td>/mæb/</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>ox</td>
<td>/ɑks/</td>
<td>1.699</td>
<td>6</td>
</tr>
<tr>
<td>pot</td>
<td>/pæt/</td>
<td>2.4472</td>
<td>34</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>rock</td>
<td>/rʌk/</td>
<td>2.8751</td>
<td>23</td>
</tr>
<tr>
<td>rot</td>
<td>/rət/</td>
<td>1.9031</td>
<td>29</td>
</tr>
<tr>
<td>shock</td>
<td>/ʃɔk/</td>
<td>2.4914</td>
<td>21</td>
</tr>
<tr>
<td>shop</td>
<td>/ʃɑp/</td>
<td>2.7993</td>
<td>16</td>
</tr>
<tr>
<td>shot</td>
<td>/ʃɔt/</td>
<td>3.0492</td>
<td>23</td>
</tr>
<tr>
<td>squad</td>
<td>/skwɔd/</td>
<td>2.2553</td>
<td>5</td>
</tr>
<tr>
<td>stock</td>
<td>/stɔk/</td>
<td>3.1673</td>
<td>12</td>
</tr>
<tr>
<td>stop</td>
<td>/stɑp/</td>
<td>3.0792</td>
<td>11</td>
</tr>
<tr>
<td>swab</td>
<td>/swæb/</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>swan</td>
<td>/swɑn/</td>
<td>1.4771</td>
<td>11</td>
</tr>
<tr>
<td>swat</td>
<td>/swɑt/</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>top</td>
<td>/tɑp/</td>
<td>3.3096</td>
<td>21</td>
</tr>
<tr>
<td>yacht</td>
<td>/jæt/</td>
<td>1.6021</td>
<td>16</td>
</tr>
<tr>
<td>boot</td>
<td>/bʌt/</td>
<td>2.1139</td>
<td>32</td>
</tr>
<tr>
<td>booth</td>
<td>/bʊθ/</td>
<td>1.8451</td>
<td>12</td>
</tr>
<tr>
<td>bruise</td>
<td>/bruz/</td>
<td>1.4771</td>
<td>10</td>
</tr>
<tr>
<td>crew</td>
<td>/kru/</td>
<td>2.5563</td>
<td>19</td>
</tr>
<tr>
<td>cute</td>
<td>/kjʊt/</td>
<td>1.699</td>
<td>6</td>
</tr>
<tr>
<td>feud</td>
<td>/fjud/</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>flute</td>
<td>/flʌt/</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>food</td>
<td>/fʊd/</td>
<td>3.1673</td>
<td>11</td>
</tr>
<tr>
<td>fruit</td>
<td>/frʊt/</td>
<td>2.5441</td>
<td>7</td>
</tr>
<tr>
<td>goose</td>
<td>/ɡʊs/</td>
<td>1.6021</td>
<td>12</td>
</tr>
<tr>
<td>group</td>
<td>/ɡrʊp/</td>
<td>3.5911</td>
<td>11</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>---------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>huge</td>
<td>/hjudʒ/</td>
<td>2.7324</td>
<td>2</td>
</tr>
<tr>
<td>juice</td>
<td>/dʒus/</td>
<td>2.0414</td>
<td>13</td>
</tr>
<tr>
<td>June</td>
<td>/dʒun/</td>
<td>2.9685</td>
<td>16</td>
</tr>
<tr>
<td>loose</td>
<td>/lus/</td>
<td>2.7243</td>
<td>23</td>
</tr>
<tr>
<td>mood</td>
<td>/mud/</td>
<td>2.5682</td>
<td>17</td>
</tr>
<tr>
<td>moon</td>
<td>/mun/</td>
<td>2.7782</td>
<td>21</td>
</tr>
<tr>
<td>moose</td>
<td>/mus/</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>move</td>
<td>/muv/</td>
<td>3.233</td>
<td>8</td>
</tr>
<tr>
<td>news</td>
<td>/nuz/</td>
<td>3.0086</td>
<td>16</td>
</tr>
<tr>
<td>nude</td>
<td>/nud/</td>
<td>2.301</td>
<td>15</td>
</tr>
<tr>
<td>ooze</td>
<td>/uz/</td>
<td>1.301</td>
<td>10</td>
</tr>
<tr>
<td>proof</td>
<td>/pruʃ/</td>
<td>2.6021</td>
<td>4</td>
</tr>
<tr>
<td>root</td>
<td>/rut/</td>
<td>2.8633</td>
<td>30</td>
</tr>
<tr>
<td>shoe</td>
<td>/ʃu/</td>
<td>2.1461</td>
<td>24</td>
</tr>
<tr>
<td>shrew</td>
<td>/ʃru/</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>sleuth</td>
<td>/sluθ/</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>smooth</td>
<td>/smud/</td>
<td>2.6232</td>
<td>1</td>
</tr>
<tr>
<td>snooze</td>
<td>/snuz/</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>soothe</td>
<td>/suð/</td>
<td>1.301</td>
<td>9</td>
</tr>
<tr>
<td>stew</td>
<td>/stu/</td>
<td>1.699</td>
<td>13</td>
</tr>
<tr>
<td>suit</td>
<td>/sut/</td>
<td>2.6812</td>
<td>27</td>
</tr>
<tr>
<td>true</td>
<td>/tru/</td>
<td>3.3636</td>
<td>18</td>
</tr>
<tr>
<td>truth</td>
<td>/truθ/</td>
<td>3.1004</td>
<td>5</td>
</tr>
<tr>
<td>tube</td>
<td>/tub/</td>
<td>2.4914</td>
<td>12</td>
</tr>
<tr>
<td>Word</td>
<td>Transcription</td>
<td>Log frequency</td>
<td>Density</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>tunic</td>
<td>/tunik/</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>view</td>
<td>/vju/</td>
<td>3.2695</td>
<td>6</td>
</tr>
<tr>
<td>youth</td>
<td>/juθ/</td>
<td>2.9138</td>
<td>7</td>
</tr>
</tbody>
</table>