UNDERSTANDING ACCEPTABILITY JUDGEMENTS: GRAMMATICAL KNOWLEDGE VS. LEXICAL SEARCH

By

Darby Grachek

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Linguistics – Master of Arts

2021

ABSTRACT

UNDERSTANDING ACCEPTABILITY JUDGEMENTS: GRAMMATICAL KNOWLEDGE VS. LEXICAL SEARCH

By

Darby Grachek

In this thesis, the source of gradience in acceptability judgments is discussed (Scholes 1966) and a set of experiments is performed which attempt to attribute gradience more concretely to either phonotactic knowledge or lexical knowledge. Two phonotactic acceptability judgment tasks are implemented to better understand whether reaction time can lessen the influence of lexical information on phonotactic acceptability judgments. Following results from Fox (1984) which show weaker influence from lexical information when less response time is allowed, I hypothesize that phonotactic information should be immediately accessible for participants, but that a lexical search takes more time to perform. In turn, an acceptability judgment task which allots less response time to participants should result in less influence from lexical information in their responses. By comparing the resulting participant judgments to gradient and categorical language models, I show that lexical access is still present at early reaction times, meaning reaction time was not useful in removing the influence of lexical information from phonotactic acceptability judgments in this set of experiments. This prompts a discussion of other possible models which can feasibly be used to understand these judgments and the source of their gradience.

ACKNOWLEDGEMENTS

I owe a huge thank you to all of the faculty in the linguistics department at MSU, but I first want to extend an extra huge thank you to my advisor, Dr. Karthik Durvasula. Karthik has not only supported my ideas and interests, but also challenged me to do more than I ever imagined I could on my own. He has helped me to become a more competent linguist, and a more confident person. I feel extremely lucky that I was able learn so much from him during these past two years.

I'm also extremely grateful for the other members of my committee, Dr. Yen-Hwei Lin and Dr. Betsy Sneller, for their insightful comments and helpful criticisms of the work that went into this thesis. Thank you to Dr. Cristina Schmitt and Dr. Alan Munn for giving me my first ever opportunity to do linguistics research and inspiring me to apply to the MA program at MSU in the first place. I also want to acknowledge Dr. Deo Ngonyani for providing me with my first teaching opportunity, and Dr. Suzanne Wagner her continuous support during my time at MSU.

Thank you to the members of the Phonogroup and Child Language Acquisition Lab for listening to my ideas, and also sharing their own. I'm also indebted to the members of the Writing Accountability Group for their incredible feedback and encouragement during the writing process for this project.

Having great friends to lean on made the biggest difference when I felt overwhelmed and discouraged, and I'm forever grateful to the great friends that struggled through class projects and deadlines with me, especially Sarah Sirna, Shannon Cousins, Yongqing Ye, Daniel Greeson, Rachel Stacey, and Michaela Smith.

I'm also always grateful to my parents, Paul and Shara, and my brother, Brooks, for supporting my dreams and ideas even when they have no idea what I'm talking about.

TABLE OF CONTENTS

LIST OF TABLES v

LIST OF FIGURES vi

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 BACKGROUND 3
1 The Source of Gradient Judgements 4
2 Evidence from Speech Perception 7

CHAPTER 3 EXPERIMENT 10
3 Stimuli 10
4 Participants and Procedure 12
5 Hypotheses 13

CHAPTER 4 RESULTS 14
6 Preliminary Results 14
6.1 Comparison to Scholes (1966) 18
6.2 Analysis of Reaction Time 19
7 Primary Results 20
7.1 Introducing the Models 20
7.2 Modeling Results 23
8 Post-Hoc Testing 29

CHAPTER 5 DISCUSSION 31

CHAPTER 6 CONCLUSION 37

APPENDIX 39

BIBLIOGRAPHY 41

LIST OF TABLES

Table 1: Table of R-Squared Values for Neighborhood Density model, Gross Phonotactic model, and MaxEnt model of mean participant responses per stimulus. 23

Table 2: R-sqaured values and confidence interval ranges for Neighborhood Density model, Gross Phonotactic model, and MaxEnt model of mean participant responses per stimulus **for the non-speeded condition only**. . 28

Table 3: List of all stimuli used in this set of experiments along with their corresponding CMU Glyphs, neighborhood density scores, gross status, and previous rating provided by participants in Scholes (1966). 40

LIST OF FIGURES

Figure 1: Counts of '1' responses per stimulus across all participants. It is separated the between speeded and non-speeded condition. The y-axis shows each stimulus in IPA, and the x-axis shows the number of times all participants responded with a '1', meaning the stimulus was judged to be a 'good' possible word of English. 15

Figure 2: A scatter plot showing a comparison of the correlation between the speeded and non-speeded conditions and a perfect correlation. The red line represents a perfect correlation (a correlation between non-speeded responses and themselves) while the blue line represents the actual correlation between the two experimental conditions. 17

Figure 3: A scatter plot showing the correlation between the total 1-responses for the speeded and non-speeded conditions and the original ratings (which are also in the form of counts) from the Scholes (1966). The y-axis presents the mean ratings for the speeded trials and non-speeded trials for the present study, and the x-axis presents the ratings from Scholes (1966). 19

Figure 4: Mean reaction time was calculated for each stimulus across participants and separated by condition. The y-axis presents each stimulus in IPA, and the x-axis measures the reaction time in seconds. 20

Figure 5: A scatter plot showing the correlation between the participant mean responses and MaxEnt harmony scores between conditions. The x-axis contains the portrays the responses means and the y-axis portrays the MaxEnt harmony scores. 24

Figure 6: A scatter plot showing the correlation between the participant mean responses and neighborhood density scores between conditions. The x-axis contains the portrays the responses means and the y-axis portrays the neighborhood density scores. 25

Figure 7: A violin plot with overlaying box plots showing a comparison between the proportion of 1-responses responses and gross status for the speeded and non-speeded conditions. The 'valid' gross status label means that the initial cluster in the stimulus is present in English, and 'invalid' label means that the initial cluster in the stimulus is not present in English. The x-axis contains the gross status information and the y-axis portrays the proportion of 1-responses. 26

Figure 8: A plot showing the R-squared values for the speeded and non-speeded conditions in each of the different models, along with the confidence intervals for each R-squared value. This shows that all the R-squared values are captured within the confidence intervals for the opposing condition, meaning that there are no significant differences between conditions for any model. 27

Figure 9: A ridge plot showing the mean ratings for each individual stimulus. These ratings are aggregated across all participants and plotted next to one another in order to discern how similarly individual participants behaved to one another between conditions for each stimulus item. . . . 30

CHAPTER 1

INTRODUCTION

Native speakers have intuitions about which combinations of sounds are considered to be 'good' or 'bad' in their language. These are referred to as phonotactic judgments. Concerning these judgments, I aim to address two main questions in my thesis: (1) Where do those judgments come from; (2) More specifically, what kinds of factors are influential in making those judgments?

It has been well-established that speaker's phonotactic judgments are gradient (Scholes 1966; Bailey and Hahn 2001), although the source of that gradience has been hotly debated. The two main factors that have been proposed in the literature as potential sources of information for phonotactic acceptability judgments are probabilistic phonotactic knowledge and lexical knowledge. However, it is not clear which of these factors is responsible for the gradience observed in previous studies.

Several studies investigating the source of this gradience have found that when modeling these judgments, language models which incorporate gradience have been found to be good predictors of participant data (Albright 2009; Hayes and Wilson 2008).

Conversely, more recent studies by Gorman (2013) and Sarver (2020) show that categorical models are just as successful at producing an accurate model of speaker judgments. This raises the question, if gradient grammatical knowledge does not significantly improve a model's ability to replicate native speaker judgments, is gradience really an integral part of the grammar? It could be possible that the gradience observed in experiments from Scholes (1966) and others is the result of factors outside of the phonotactic grammar, such as lexical knowledge or task effects.

The source of gradience in phonotactic judgments is of importance for understanding the scope of what phonological theory needs to account for. In claiming that gradient phonotactic judgments are possible, we necessitate that our theory of phonology has to be able to process

fine-grained information in a gradient fashion. This is a much more complex version of the phonology compared to a the categorical view. In order to avoid this overly-complex view of phonological theory, I seek to provide evidence for categorical phonotactic judgments (Berg 2018). This would mean that the phonology does not need to account for the gradient judgments seen in previous phonotactic acceptability judgment tasks. Instead, it would be possible to attribute that gradience to an extra-phonological factor like lexical information or task effects.

In this thesis, I investigated whether the gradience observed in previous judgments from Scholes (1966) and others is really the result of the phonotactic grammar, or if influence from lexical knowledge is actually responsible. In order to tease apart these two factors, I preformed a set of experiments inspired by Fox (1984), who studied the influence of lexical knowledge on the Ganong effect by looking at the effect of reaction time on participant responses. The intent of this set of experiments was to separate the influence of lexical knowledge from probabilistic phonotactic knowledge in speaker's judgments, in order to attribute the gradience in speaker judgments clearly to either phonotactic knowledge or lexical knowledge. Results show that differences in response time are not enough to separate these two factors prompting a discussion of the Cohort Model and other ways to determine the role the lexical and phonotactic knowledge in phonotactic acceptability judgements.

CHAPTER 2

BACKGROUND

An early theory of phonotactic knowledge and judgments was proposed by Chomsky and Halle (1968). They observed that speakers have clear categorical preferences for certain sound sequences over others, even in novel words. For example, there is a clear difference in the acceptability of the two novel words [blɪk] and [bnɪk]. Even though both are nonce words and are therefore unattested as words of English, [blɪk] is judged to be acceptable while [bnɪk] is not.

Chomsky and Halle's (1968) explanation for that preference is the existence of sequence structure constraints. These are feature-based constraints on which types of segments can occur next to each other in a sound sequence. Given the judgments of native speakers, preferences for some sequences over others can be modeled through a constraint against word initial stop and nasal combinations, with no such constraint against stop and liquid combinations.

Contrasting with Chomsky and Halle's categorical account, evidence of gradient phonotactic judgments has also been attested. Scholes (1966) investigated whether speakers could assign different 'levels' of acceptability, showing that speaker judgments are gradient, not categorical. In order to demonstrate this gradience, he conducted an experiment where novel word stimuli were presented auditorily to 35 seventh-graders with the following prompt:

> Suppose these are foreign words which English wishes to borrow; which ones will be admitted in their present form and which ones will be changed?

These same participants were also told, falsely, that some of the words they were hearing were real words of English that the participants simply had not heard of before. This was done presumably to encourage speakers to treat the nonce words more like existing English

words and better ensure that nonce word judgments would be comparable to those of real English words.

Stimuli were all in the form of CCVC(C), with the initial consonant cluster being used to test speaker's judgments on phonotactic grammaticality. The participants heard a stimulus, and responded with either 'yes' (it could be borrowed into English) or 'no' (it could not be borrowed into English) on a worksheet.

Results showed that participant responses varied greatly and indicated that there are in fact different 'levels' of acceptability in participant's judgments of phonotactic acceptability. Scholes concluded from this that any model which seeks to replicate speaker phonotactic acceptability judgments would have to take into account that participants seem to have a gradient measure of phonotactic acceptability.

1 The Source of Gradient Judgements

While a number of works add empirical support to Scholes' (1966) claims that phonotactic judgments are gradient (Albright 2007; Albright and Hayes 2003; Bailey and Hahn 2001), the *source* of that gradience is often disagreed upon. Some claim that lexical information is responsible for the observed gradient judgments, while others claim that probabilistic phonotactic information is actually the source. For those that claim lexical information is the source, factors like neighborhood density are often cited as being correlated with acceptability judgments. Neighborhood density is the number of words that are similar to the target word, usually by a one phoneme difference. Vitevitch and Luce (1999) found that neighborhood density has an effect on how quickly participants are able to respond to word and non-word stimuli. Specifically, high neighborhood density words are perceived slower and less accurately than low neighborhood density words, but novel words with high neighborhood density are responded to more quickly than low density novel words.

Bailey and Hahn (2001) found that probabilistic phonotactic information and lexical information (namely, neighborhood density and token frequency) both have a significant

effect on speaker judgments. They found that lexical information accounted for 23% of the variance in speaker judgments (or 29% using their Generalized Neighborhood Model which also takes frequency information into account), while the model of probabilistic phonotactic information they tested accounted for 18% of the variance. While this shows that both factors could affect speaker's judgments about what is a possible word of English, it also shows that there is a very large portion of the variance in speaker judgments that is unaccounted for by both factors. This could either mean that there are factors besides probabilistic phonotactic knowledge and lexical knowledge that affect the level of gradience in speaker judgments, or that the current models in use lack the nuance to correctly identify the source of more variance in the results.

Albright (2007) claims that a phonotactic grammar, defined as a grammar that includes bigram probabilities, is the main source of gradient judgments. In order to show this, he used phonotactic probabilities to create a model of gradient participant behavior. He claims that using these probabilities to model participant judgments does a better job of replicating gradient responses in phonotactic judgment tasks than those which use lexical information like neighborhood density.

Shademan (2006) also suggests that probabilistic phonotactic information is a better predictor of speaker judgments. She investigated the possibility that variance in speaker judgments could be a task effect related to whether or not there are real words in the stimuli. In order to demonstrate this, she implemented two versions of a nonce word acceptability judgment task. One condition contained only nonce words, and the other contained both nonce words and real words. Stimuli varied in lexical similarity using neighborhood density (low and high) and in phonotactic probability (high probability, low probability, and phonotactic violations were all present). For the acceptability judgment tasks, Shademan observed phonotactic probability ratings were similar across both conditions. That is, participants consistently rated the lower probability forms as being less acceptable than the higher probability forms (as expected) in both conditions. However, the effect of lexical density was

not consistent between the two conditions. Shademan found that participants who saw the condition with both real and nonce words gave slightly lower ratings to the nonce words with low neighborhood density than those who saw only the nonce word condition. This suggests that the effects of lexical similarity may be more sensitive to task effects like whether or not real words are present in the stimuli. This, Shademan claims, makes phonotactic probability more stable and therefore, a better predictor of speaker judgments.

However, it might also be the case that phonotactic information is not simply less sensitive to task effects (and more reliable for producing speaker judgments as a result), but it is also not the source of the variability in speaker judgment. If there is more variability in speaker judgments with regard to lexical information, this makes a better case for lexical information being the source of gradient judgments, instead of phonotactic probabilities.

While several studies have found probabilistic phonotactic knowledge to be a more reliable factor in capturing the gradient behavior of phonotactic acceptability judgments (Albright 2007; Bailey and Hahn 2001; Shademan 2006), there is also evidence that lexical knowledge is equally accurate in replicating these judgments. One such study, conducted by Gorman (2013), compares 4 different models to measure which model types most accurately account for speaker judgments. The first model, the Gross Phonotactic Model, evaluated data by categorizing each nonce word as either well-formed (containing no phonotactic violations), or ill-formed (containing phonotactic violations). A token is considered well-formed when the constituent sequences have a non-zero frequency in a representative sample.[1] This is evaluated at both the level of the onset and the rhyme.

The second model used neighborhood density as a measure to assess whether a nonce word would be judged favorably by participants. Gorman adopts the one-phoneme metric to define neighborhood density, meaning that a word which can be formed by making a one-phoneme change to the nonce word in question is counted as a 'neighbor'. The more lexical neighbors a word has, the more likely it is to be rated as phonotactically acceptable.

[1]Refers to whether the sound sequence is present in the Carnegie Mellon University dictionary (Weide 1994).

The third model measured word-likeliness by using bigram frequency, and the fourth model is the MaxEnt model, which claims that phonotactic judgements can be modeled by using assigned weights according to the principle of maximum entropy (Hayes and Wilson 2008). This model can supposedly capture both categorical and gradient phonotactic patterns and does not need to be provided with constraints in advance. Instead, it only needs to be trained on a sample set of words.

Gorman found that the Gross Phonotactic Model (a categorical model) and the Neighborhood Density Model performed better than both the bigram frequency model and the MaxEnt model. That is, they are able to more reliably predict participant acceptability judgments. This shows that gradient models do not reliably predict intermediate ratings for nonce words. Importantly, Gorman also found that the Gross Phonotactic Model (which uses phonotactic knowledge) and the Neighborhood Density Model (which uses lexical information) perform at about the same level.

2 Evidence from Speech Perception

The above discussion clearly shows that there are multiple different views on the source of the gradience in speaker judgments. Adding to this is the fact that experimental results will always look somewhat gradient when comparing the behavior of many independent participant judgments (Armstrong, L. R. Gleitman, and H. Gleitman 1983). Clearly, it is important to find some way to be objective in measuring gradience. One way to forge ahead is to see how other related domains have dealt with the issue of multiple sources. In this thesis, I will use a technique inspired by work from the speech perception literature that attempts to tease apart lexical knowledge from early perception.

A specific case in speech perception where a similar question arose is the Ganong effect. The Ganong effect is a phenomenon in which an ambiguous segment is more likely to be identified as one sound over another if the sound forms a word with the surrounding segments. For example, a sound that is ambiguous between [t] and [d] might be more often identified

as a [d] when it is heard in the context of the rhyme [-æʃ], since *tash* isn't a word of English (and *dash* is). However, if the ambiguous segment was heard in the context of the rhyme [-ɛkst], participants would be more likely to identify the first segment as a [t], since *text* is a real English word and *dext* is not. By implementing speech perception research, Fox (1984) developed a better understanding of the Ganong effect and its interaction with lexical knowledge. His goal was to investigate the difference between mechanisms of the speech perception system, and post-perceptual decision-making mechanisms, which can utilize such factors as lexical information, using reaction time to mediate between them.

Previous accounts of the Ganong effect characterized its interaction with the speech perception process by using the criterion-shift model (Ganong 1980). This model allows lexical information to affect and bias the process of phonetic categorization, meaning that it occurs simultaneously with phonetic perception. However, Fox claims that the difference in responses is a product of a post-categorization process where possible phonetic characterizations for each of the non-word tokens would be changed to categorizations that formed real words after initial phonetic categorizations were made. This would mean that phonetic categorization occurs first, and afterwards, lexical information influences those categorizations. This model is referred to as the categorical model.

In order to provide evidence for the categorical model, Fox set up two experiments with segments that were ambiguous between [b] and [d]. Some of the ambiguous segments were onsets for real English words, and some were onsets of nonce words. Participants were then asked to identify whether the first segment was a [b] or a [d]. The only difference between the two experiments was that in one version, participants were told to respond as quickly as possible, and in the other, they were not given any time limit on their responses.

Results showed that the more time participants had to provide their response, the more they responded with the segment that formed a real English word. This suggests that when participants are given less time to process a stimulus, there is less influence from lexical knowledge. This also suggests that the categorical model provides a better description of

the speech perception process.

More recent work has questioned this result (Rysling et al. 2015; Kingston et al. 2016) showing that lexical knowledge is not completely absent at shorter reaction times, but it is instead lessened compared to longer reaction times. This is somewhat concerning for Fox's findings, but it is at least still true that lexical knowledge has less influence over participant decisions at shorter reaction times.

Fox's findings lead us to question whether the influence of lexical knowledge on phonotactic acceptability judgments can be separated to some extent from probabilistic phonotactic knowledge using reaction time as a way to mediate its presence.

In this thesis I will use methodology outlined by Fox (1984) in order to further our understanding of phonotactic acceptability judgments. Specifically, how does timing affect the influence of lexical knowledge on phonotactic acceptability judgments, and is there a reduction in the gradience of said judgments when participants are given less time to make a decision?

Taking Fox's line of reasoning into account, it is possible that pushing participants to answer quickly would allow less time for post-perceptual mechanisms, like lexical knowledge, to influence their decision. If in fact the gradience in judgments comes from some sort of lexical search or lexical comparison, that should be easier to conduct with additional time. Therefore, the acceptability judgments should become more gradient (in line with lexical statistics) as the participant has more time.

CHAPTER 3

EXPERIMENT

In order to examine the effect of reaction time on phonotactic acceptability judgments, I performed two phonotactic acceptability judgment tasks - one that is speeded, and one that is not. If the gradience in the responses comes largely from a lexical comparison/search, then the speeded task should not allow for as much gradience in participant responses as the non-speeded task. This is because there is less time available for a search of the lexicon to be conducted and therefore less lexical knowledge is available to influence acceptability judgments.

3 Stimuli

Stimuli included 46 of the nonsense words from Scholes' (1966) study, selected from the original 61 so that there is still a range of licit to illicit phonotactic sequences, plus four additional stimuli added in order to incorporate more possible clusters like [θw-] (full list of stimuli located in the appendix). All stimuli are monosyllabic nonce words and contain a number of different bi-consonantal clusters. There were 50 total clusters ranging from those that occur in English ([sm-], [dr-]), to clusters that never occur in English ([zf-], [bv-]). The stimuli also incorporated a number of different rhymes so that participants won't be clued in to the fact that the onset clusters are what is being tested. Each stimulus was presented four times in order to capture gradience at the individual level for each participant. An equal ratio of fillers to stimuli were also included, which were also repeated four times each. This was done in order to ensure that participants remain sufficiently ignorant to the task's focus on the consonant clusters in the stimuli. This means there were 50 stimuli and 50 fillers which each repeated four times resulting in a total of 400 tokens for each participant. All stimuli were recorded by a native American English speaker (the author) using a Samson USB Studio Meteor Microphone.

Admittedly, such a large number of stimuli made the experiment quite long and potentially exhausting for participants. It is true that it is harder to trust judgments from speakers who are exhausted, especially when the experiment is administered online as this experiment was. To combat this, I incorporated an optional two-minute break into the middle of the experiment (after the first 200 tokens were presented) so that participants were able to rest during the experiment if necessary. If a participant wanted to end their break early, they could press the spacebar to skip the break and continue with the task.

As for fillers, careful thought was put into the nonce words used as fillers in this experiment. This is because using clusters that are more complex than the clusters used in the test stimuli (e.g. tri-consonantal clusters like [ftl-]), might have resulted in skewed judgments. In other words, having very 'bad' clusters in the fillers might have pushed some of the judgments of the test clusters, which are less common or not present in English, farther in the direction of being 'good', simply because they are not as shockingly bad as the tri-consonantal cluster in the example above. In order to avoid skewing judgments, the fillers contained mostly simplex onsets. If they did have complex consonant clusters, they were not in the onset of the word (to contrast them from the stimuli), but in the coda instead. This was done to draw the participant's attention away from the clusters when making judgments about words.

Another aspect of the stimuli that required careful consideration was the process of recording the clusters for each experimental token. Having unattested consonant clusters in the stimuli means that there may be small articulatory differences in the way that a native English speaker pronounces clusters that are present in English, and those that are not. For example, an English speaker may inadvertently produce a schwa vowel between each segment in the cluster [bd-]. This might lead participants to judge those tokens, not as clusters, but as CVCVC sequences ([bəda] instead of [bda]). In order to avoid this, I produced all of the stimuli with a schwa vowel between the first and second segment of the bi-consonantal clusters. Then, that schwa vowel was spliced out using Praat (Boersma and Weenink 2016) so that there were no accidental articulatory cues in the speech signal. This was done by

splicing each token at zero-crossings in order to make the stimuli sound as close to natural speech as possible. For example, to get the test stimulus [bna], I pronounced [bəna], then spliced out the [ə] between the [b] and [n] segments.

4 Participants and Procedure

One-hundred-thirteen participants were recruited online via Prolific (Palanab and Schitter 2018), but thirty-four were disqualified as they did not finish the experiment. Seventy-nine participants completed the study and were paid $10.11 per hour for their participation. Most participants took around twenty minutes to complete the entire experiment, meaning payment for each person was around $3.37. However, nineteen additional participants were eliminated for either not responding to any stimuli, or having answers which were too uniform showing that the participant was just clicking through the task without paying proper attention. Ultimately, sixty participants were used for data analysis. A between-subjects design was implemented to run the two different versions of the experiment, meaning that thirty participants received the speeded acceptability judgement task, and thirty received the non-speeded task. This was done in order to make sure that participants did not have any previous information about the stimuli used in each task (since the same stimuli will be used in both). PsychoPy (Peirce et al. 2019) was used to design the experiment, and and Pavlovia (Peirce et al. 2019; Palanab and Schitter 2018) was used to run the experiment online.

During the experiment, participants were presented with each stimulus auditorily and asked to respond whether they thought the nonce word was 'good' or 'bad' as an example of a typical word of English. The crucial experiment design manipulation was that one version allowed participants to respond at their own pace, and the other pushed participants to respond as quickly as possible. This was implemented via a training session where participants were encouraged to answer as quickly as possible during the speeded task. They were also told that if they did not answer quickly enough, the experiment would move on without

them and their response would not be recorded. This was not present in the non-speeded trial. Instead, the experiment would not continue until it received the participants answer. The inter-trial interval (ITI) for both tasks was 1000 ms.

During each trial, participants were given the prompt "What do you think of the following word as a possible word of English?" and each nonce word stimulus was presented one at a time. Participants were then asked to provide their acceptability judgments for the stimulus, which they selected by pressing either the '1' key for good, or the '0' key for bad.

I chose to utilize a binary choice component in this set of experiments for two reasons: One is to replicate Scholes' original 1966 experiment where he also only allowed a binary response. The other is that I assumed that in a speeded task like the one used here, that a Likert scale would not be fully explored by participants due to the speed with which they were meant to respond to the stimuli. In order to balance the two experiments, I used a binary forced choice design for both the speeded and non-speeded version of the task.

5 Hypotheses

Based on Fox (1984), I hypothesized that in the speeded trials, participant's judgments would be less gradient than in the non-speeded trials. Assuming that the gradience in a speaker's judgment comes from lexical information such as neighborhood density or frequency, it should take more time for the speaker to perform a search of the lexicon than it should for them to rely solely on phonotactic knowledge. In the speeded trials, I hypothesized that participant judgments would become more dependent on phonotactic knowledge. If Shademan (2006) is correct, then the responses using phonotactic knowledge alone should be less varied than those using lexical knowledge. Whether or not lexical information is being used will be dictated by how much time the participant has to make their judgment. In other words, the speeded trials should prevent lexical information from being used, prompting a less gradient response than the non-speeded trials (where lexical information is readily available).

CHAPTER 4

RESULTS

All data analysis was conducted in R (R Core Team 2019) using the meta-library `tidyverse` (Wickham et al. 2019). The results are split up into two sections: the preliminary analysis and the primary analysis. The preliminary analysis was preformed in order to have a more holistic view of the data. These preliminary results include a comparison of the number of '1' responses ('good' responses) in each condition, a discussion of regression towards the mean, correlations between conditions, an analysis of reaction time, and a comparison between the present study and a previous phonotactic acceptability judgement task by Scholes (1966).

The primary results contain model comparisons which were used to evaluate whether the patterns of responses between the speeded and non-speeded conditions were different from one another in a more objective fashion, which is central to the assessment of my hypotheses.

6 Preliminary Results

In order to understand how participants responded to each stimulus, the total number of '1' responses (meaning the participants thought the nonce word sounded like English) per stimulus were analyzed according to experimental condition (speeded and non-speeded). Figure 1 below shows that participant responses between conditions were quite similar.[1] Many of the highest rated stimuli in the speeded condition are also some of the highest rated stimuli for the non-speeded condition and vice versa. Figure 1 also contains information about the gross status of the stimuli, which in this case is a measure of whether the initial

[1] A note about the scale of the x-axis: This scale is in terms of raw counts - it goes up to around 60-70 because each stimulus was repeated 4 times for 30 participants. This means that for each stimulus, there are 120 total times that a stimulus can be responded to by the participants for each condition. Looking at proportions of participant responses, they responded with '0', meaning they thought the nonce word would not be a good example of a typical word of English, about half the time. This checks out when observing the totals for the counts of '1' responses because 60 is about half of 120.

cluster is observed in English. This is calculated by verifying whether the cluster is present in a representative sample of English (here the Carnegie Mellon University Pronunciation dictionary is used (Weide 1994)). If a stimulus is labeled as 'valid', that means that its initial cluster is present in English, and if a stimulus is labeled as 'invalid', that means its initial cluster is not present in English.

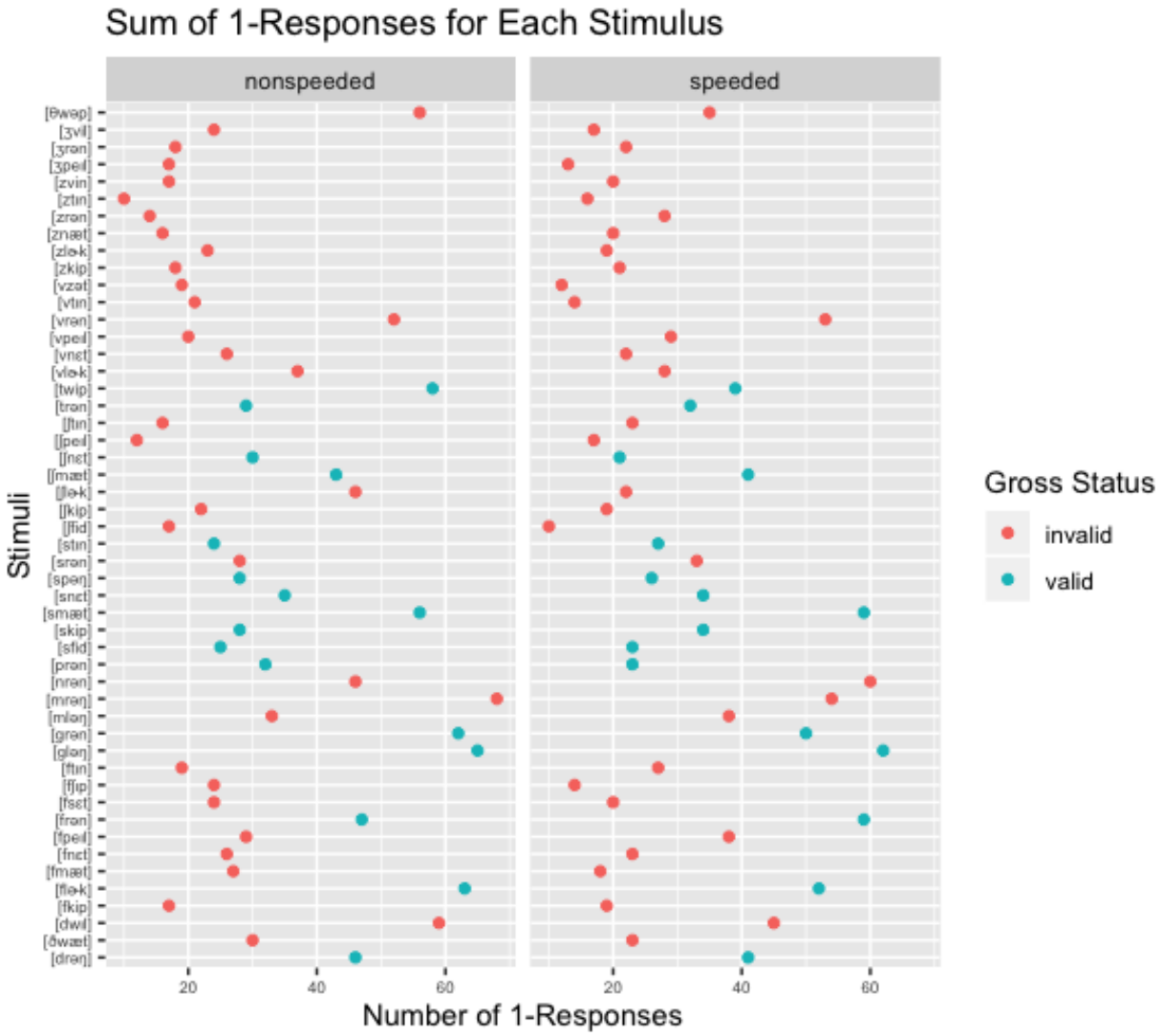

Figure 1: Counts of '1' responses per stimulus across all participants. It is separated the between speeded and non-speeded condition. The y-axis shows each stimulus in IPA, and the x-axis shows the number of times all participants responded with a '1', meaning the stimulus was judged to be a 'good' possible word of English.

Figure 1 also clearly shows that the responses for the speeded condition tend to be closer

to the center of the x-axis than those in the non-speeded condition. This can be interpreted as regression toward the mean. (This will be more salient in Figure 2). A regression toward the mean indicates that participants in the speeded trial were answering with responses that were closer to the mean of the data than in the non-speeded version. This could be interpreted in two ways: In one case, regression toward the mean would be interpreted as there being less variation in participant responses for this condition (as my hypothesis predicts). Another interpretation is that participants in the speeded condition were just guessing more often than the non-speeded participants. Of the two possible interpretations, the latter is more detrimental to the interests of the present study because it could allude to there being a less systematic response strategy for the participants in the speeded condition. Under that assumption, it may be that since participants were under more pressure to answer quickly, they were more likely to randomly guess at an answer rather than produce a judgement rooted in phonotactic knowledge.

One way to challenge the notion that participants were guessing more often in the speeded condition is to compare how similar the standard deviation is for responses to each stimulus between the two conditions. The standard deviations of the responses for both conditions are quite similar to one another (speeded = 14.18, non-speeded = 15.94; standard deviation was calculated using the counts for each stimulus aggregated across participants). It is worth noting that the standard deviation is slightly lower for the speeded condition than the non-speeded condition (this is the right direction for my prediction that the speeded version would be less variable than the non-speeded due to less influence from the lexicon). However, the similar standard deviation scores suggest that the answers given by participants for each condition were similar enough to rule out the second interpretation of the speeded responses' regression toward the mean - that the participants were just randomly guessing more often for the speeded task. This can be further verified by considering the correlation between the responses for each condition.

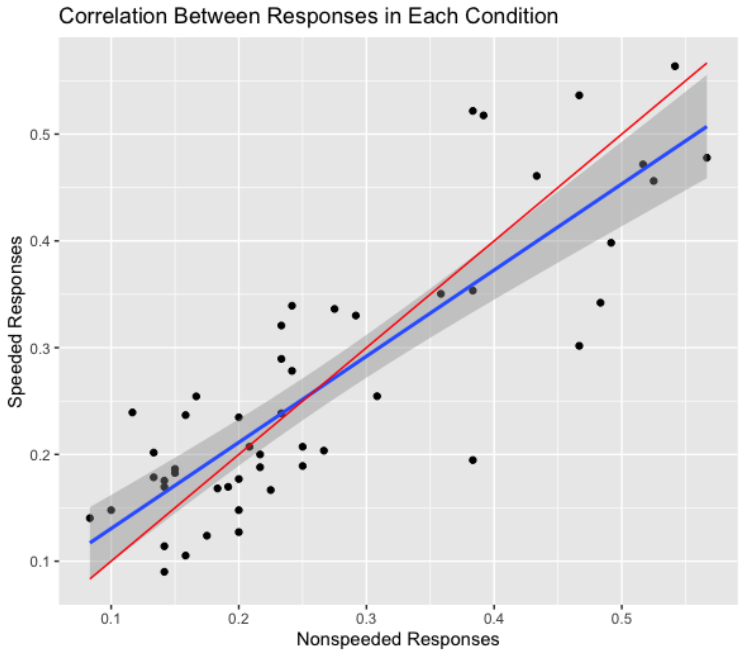

Figure 2: A scatter plot showing a comparison of the correlation between the speeded and non-speeded conditions and a perfect correlation. The red line represents a perfect correlation (a correlation between non-speeded responses and themselves) while the blue line represents the actual correlation between the two experimental conditions.

The correlation value between the two conditions is 0.85 (quite high) which can also be visualized in Figure 2 above. The red line portrays a perfect correlation (in this case, it is a correlation between the non-speeded condition and itself), and the blue line is the correlation between the speeded and non-speeded conditions. If the blue line is close to the red line, it would suggest that the results of the two experiments are very similar. The red and blue lines are in fact quite close, providing evidence for highly correlated responses between conditions. This shows that the speeded and non-speeded responses are actually highly correlated, which makes the possibility of the speeded responses being the result of random guessing less probable.

However, does this mean that the responses for each condition are so similar that the previously reported difference in the level of variance is not significant? One way to answer this question is to look at the confidence intervals of the correlation value. The null hypothesis predicts that there is no difference between the pattern of responses in each condition,

meaning that the correlation value should be 1. If the confidence interval contains the value 1, that would suggest my result as not significantly different from the prediction made by the null hypothesis. The actual confidence interval for the correlation value is 0.7416 - 0.9097, suggesting that there is some evidence pointing to participants having substantially different behavior between the two experimental conditions. Although participant response patterns are highly correlated, they are potentially different enough to provide evidence for a different response pattern between the speeded and non-speeded tasks. However, a correlation may not be the best way to look at a difference in the level of gradience between conditions. In order to be objective about how gradience is measured, I also fitted the data of both conditions to gradient and categorical models, which is outlined in section 4.2.2.

6.1 Comparison to Scholes (1966)

It is also a useful check to compare the results of this experiment to the results of Scholes (1966), since Scholes had a similar procedure to the one used here, and a subset of the stimuli used in Scholes (1966) were used in the present study. The correlation between the participants responses for the non-speeded condition is 0.69, and 0.63 for the speeded condition. It makes sense that the non-speeded version has a slightly higher correlation than the speeded one since Scholes' task did not involve a time limit for responses. The correlation value for the non-speeded version suggests a positive correlation between between the behavior of participants in both studies. A scatter plot that shows evidence of the correlation between the responses from Scholes (1966) and both the speeded and non-speeded conditions is shown below in Figure 3.

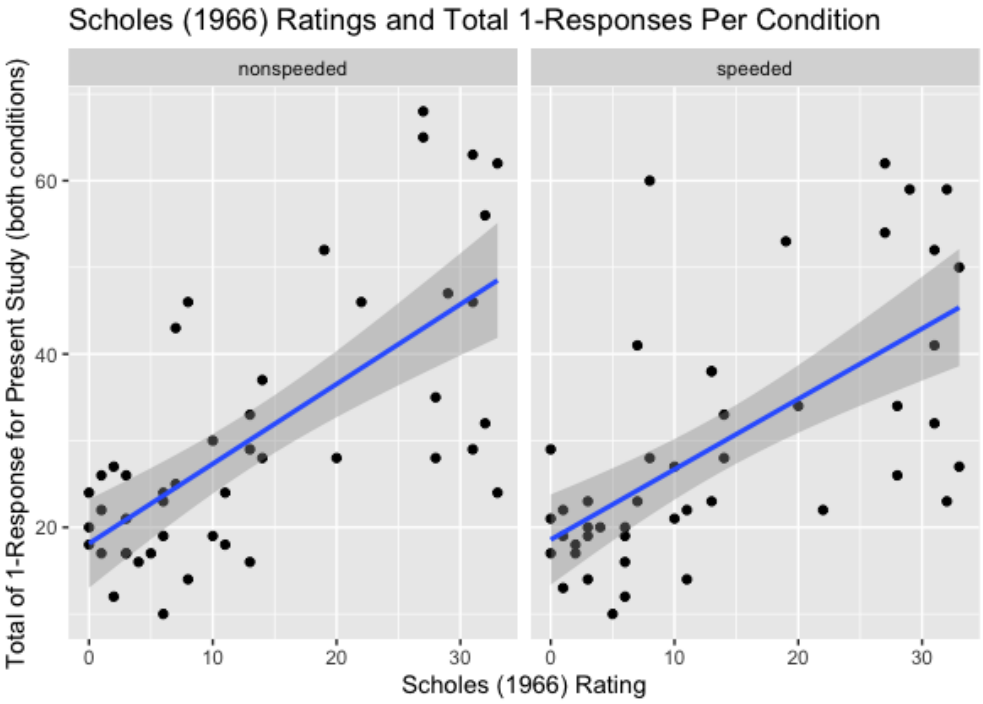

Figure 3: A scatter plot showing the correlation between the total 1-responses for the speeded and non-speeded conditions and the original ratings (which are also in the form of counts) from the Scholes (1966). The y-axis presents the mean ratings for the speeded trials and non-speeded trials for the present study, and the x-axis presents the ratings from Scholes (1966).

6.2 Analysis of Reaction Time

An analysis of the reaction times between the speeded and non-speeded conditions will inform us about how well the experimental designs for each condition were able to regulate the speed at which participants responded. The speeded condition timed out and did not record a participant's answer if they were too slow with their response. This resulted in a loss of 695 responses in the speeded condition (this is equal to .05% of the responses from each participant). Ideally, the responses for the speeded condition will be consistently faster than the responses in the non-speeded condition.

As expected, the reaction times in the non-speeded condition are much more varied than in the speeded condition. The average reaction time for the speeded trials is much shorter at 934 ms than the average reaction time in the non-speeded trials at 1388 ms (a difference

of about 450 ms). This can also be seen in Figure 4.

A correlation between the speeded and non-speeded reaction times shows that the two are not highly correlated (0.24). A weak correlation between the reaction times for both conditions shows that the mean of reaction time between the speeded and non-speeded conditions was significantly different.

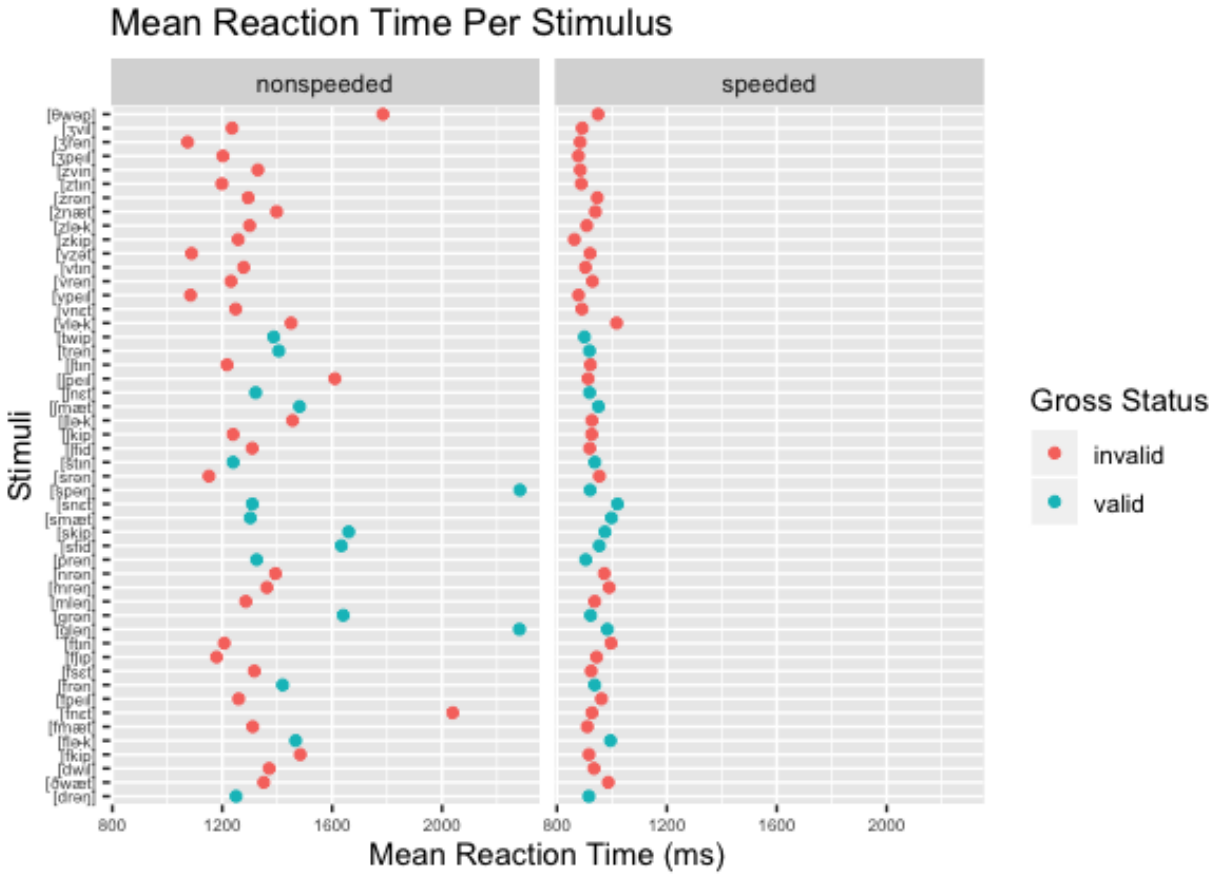

Figure 4: Mean reaction time was calculated for each stimulus across participants and separated by condition. The y-axis presents each stimulus in IPA, and the x-axis measures the reaction time in seconds.

7 Primary Results

7.1 Introducing the Models

It is important to consider how gradience should be measured when observing the pattern of participant responses between each condition. It is true that random variation is

always possible in experiments as a result of noise in participant behavior (Armstrong, L. R. Gleitman, and H. Gleitman 1983). Because of this, it might be difficult to tease apart genuine gradience in phonotactic judgments from inevitable non-systematic variation between different participants. In order to separate systematic gradience from non-systematic variation, model comparisons were performed between the mean of participant responses for each stimulus and several different types of models. The goal of these model comparisons was to measure differences between the response patterns in the speeded and non-speeded conditions more objectively (i.e. avoiding incorporating non-systematic variation into the comparison).

In order to perform each model comparison, mean participant responses per stimulus item were fitted to several models of phonotactic acceptability. My original hypothesis was that there is a decrease in the influence of lexical knowledge at shorter reaction times. If lexical knowledge is hypothesized to be responsible for the gradience in participant judgements, then the non-speeded task should have more gradience in the responses than the speeded task. This is because participants will have had more time to access lexical knowledge than in the speeded task. In order to find evidence for that claim, the results of the speeded condition and non-speeded condition would need to show significantly different effect sizes for each of the models. If the model fits decrease between the non-speeded and speeded conditions, it may be possible to reason that the relevant knowledge that the model is trying to capture is used less in the speeded condition. If it is true that lexical knowledge is used less in the speeded condition, and the gradience is largely coming from such knowledge, then the fits of the gradient models should worsen in the speeded task. In contrast, the fit of the categorical model should improve in the speeded condition.

The three specific models that were used in this analysis are the MaxEnt Model (Hayes and Wilson 2008), the Neighborhood Density Model, and the Gross Phonotactic Model (Gorman 2013).

The first model mentioned above, the MaxEnt Model (Hayes and Wilson 2008), is a

gradient model that uses weighted constraints pre-determined by a training data set to establish the acceptability of a given word. The model calculates a harmony score for each word input into the model using the following equation:

$$h(x) = \sum_{i=1}^{N} W_i C_i(x) \tag{4.1}$$

Here, the weight of the constraint i is represented by W_i, while C_i represents the number of violations of that constraint, and the summation over all of the constraints is represented by $\sum_{i=1}^{N}$. The harmony score of x (h(x)) is calculated by summing the products of the weight of each constraint and the number of violations of the corresponding constraint. These scores were obtained via the UCLA phonotactic learner (Hayes and Wilson 2008). The MaxEnt model uses only phonological information to form the harmony scores that will be used to estimate participant acceptability judgements. If the data fits best to this model, it would show that a grammar which is able to use gradient phonotactic information when making phonotactic judgements is the most accurate way to conceptualize the patterns in the data.

The second model implemented here is the Neighborhood Density Model, which uses neighborhood density scores as a way to rate words as acceptable or unacceptable. These scores were collected using the iPhod neighborhood density calculator (Vaden, Halpin, and Hickok 2009). The higher the neighborhood density score, the more acceptable the word should be for participants. This is an important model for us to compare participant responses to because it uses lexical information to make its predictions about phonotactic acceptability. If the data fits better to this model than other models, that would mean that a lexical factor is better at explaining the patterns in the data than a phonotactic one.

The third model, the Gross Phonotactic Model, is a categorical model that uses gross phonotactic information to judge whether words are acceptable or not. In this case, the gross status of a cluster is considered valid if it is present in the Carnegie Mellon University dictionary (Weide 1994), and invalid if is not present. Essentially, a cluster is considered 'good' if it occurs in English and 'bad' if it does not. If the data fits best to this model, it

would show that a categorical measure utilizing only phonotactic information does the best job of explaining the patterns in participant responses.

In order to compare the two experimental conditions to neighborhood density and gross phonotactic status, a linear regression was performed between the by-item average participant response (for each condition) and the neighborhood density scores for each stimulus. A separate regression was performed between mean participant response and gross status for each stimulus. A regression analysis was also performed between mean participant response and the MaxEnt harmony scores for each stimulus. The central comparison for my hypothesis is between the difference in the effect size of speeded and non-speeded conditions for each model. A general analysis of which model has the best fit to the data is also performed in order to compare these results to the results of previous studies, but remains secondary to the analysis of the model fits between conditions.

7.2 Modeling Results

The resulting R-squared values from each of the models are recorded in Table 1 below. This shows the effect size of the comparison between the model and the responses (basically how well the responses fit each particular model). The crucial comparison is between the speeded and non-speeded effect sizes for each model, with a higher R-squared value indicating a better fit to the model. Additionally, Figure 5 shows the correlations between participant response means and the MaxEnt harmony scores, while Figure 6 shows the correlation between the participant response means and neighborhood density. Figure 7 shows a comparison in the proportion of 1-responses and gross status. These plots all show that in each case, the distribution of responses between conditions is quite similar.

	Neighborhood Density	Gross Status	MaxEnt
Speeded	0.19	0.21	0.39
Non-speeded	0.10	0.19	0.35

Table 1: Table of R-Squared Values for Neighborhood Density model, Gross Phonotactic model, and MaxEnt model of mean participant responses per stimulus.

Figure 5: A scatter plot showing the correlation between the participant mean responses and MaxEnt harmony scores between conditions. The x-axis contains the portrays the responses means and the y-axis portrays the MaxEnt harmony scores.

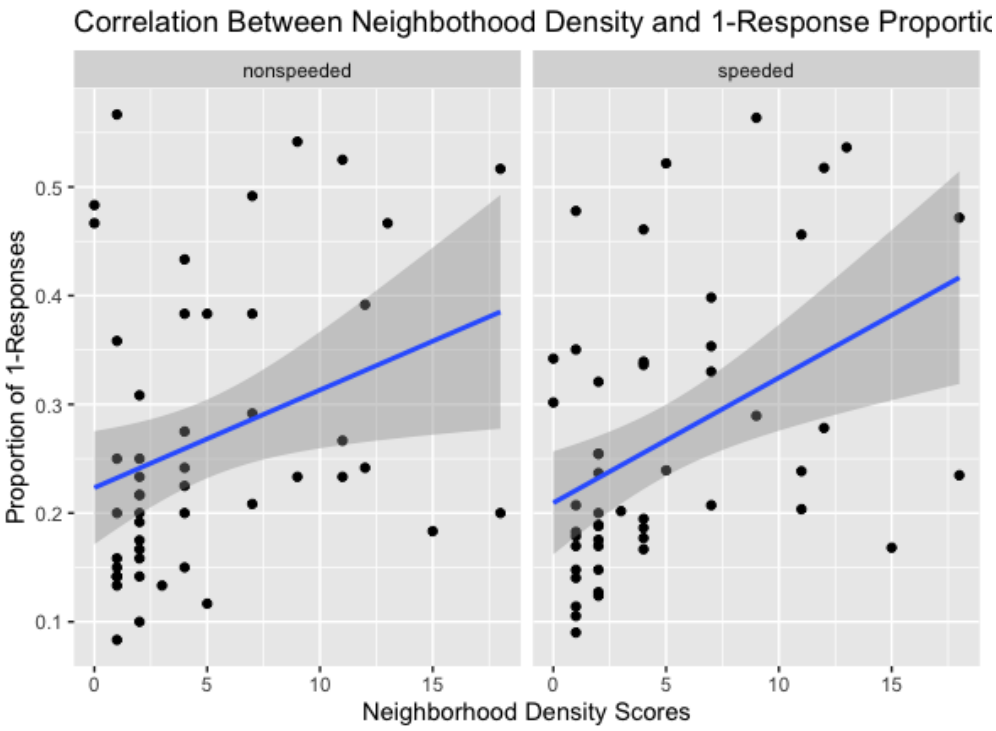

Figure 6: A scatter plot showing the correlation between the participant mean responses and neighborhood density scores between conditions. The x-axis contains the portrays the responses means and the y-axis portrays the neighborhood density scores.

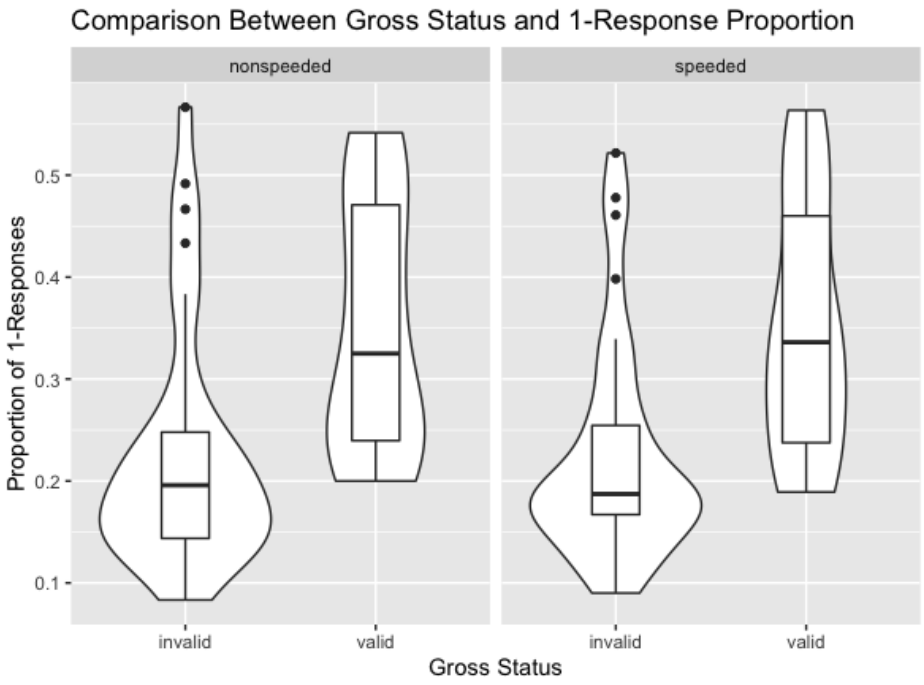

Figure 7: A violin plot with overlaying box plots showing a comparison between the proportion of 1-responses responses and gross status for the speeded and non-speeded conditions. The 'valid' gross status label means that the initial cluster in the stimulus is present in English, and 'invalid' label means that the initial cluster in the stimulus is not present in English. The x-axis contains the gross status information and the y-axis portrays the proportion of 1-responses.

The next step for this model comparison is finding out whether there are significant differences between the model effect sizes for the speeded and non-speeded conditions. Once again, confidence intervals were calculated for each model to see if the R-squared value for one condition is contained within the confidence interval of the other condition. If the confidence intervals for the speeded conditions for each model do not contain the R-squared value for the non-speeded conditions (and vice versa), one can conclude that the differences in the R-squared values (and therefore the model effect sizes) are significantly different between conditions. This would provide evidence for participant acceptability judgements being significantly affected by the amount of time they were given to respond. The confidence intervals for each model are visible in Figure 8 below as the bars surrounding the R-squared values for each model:

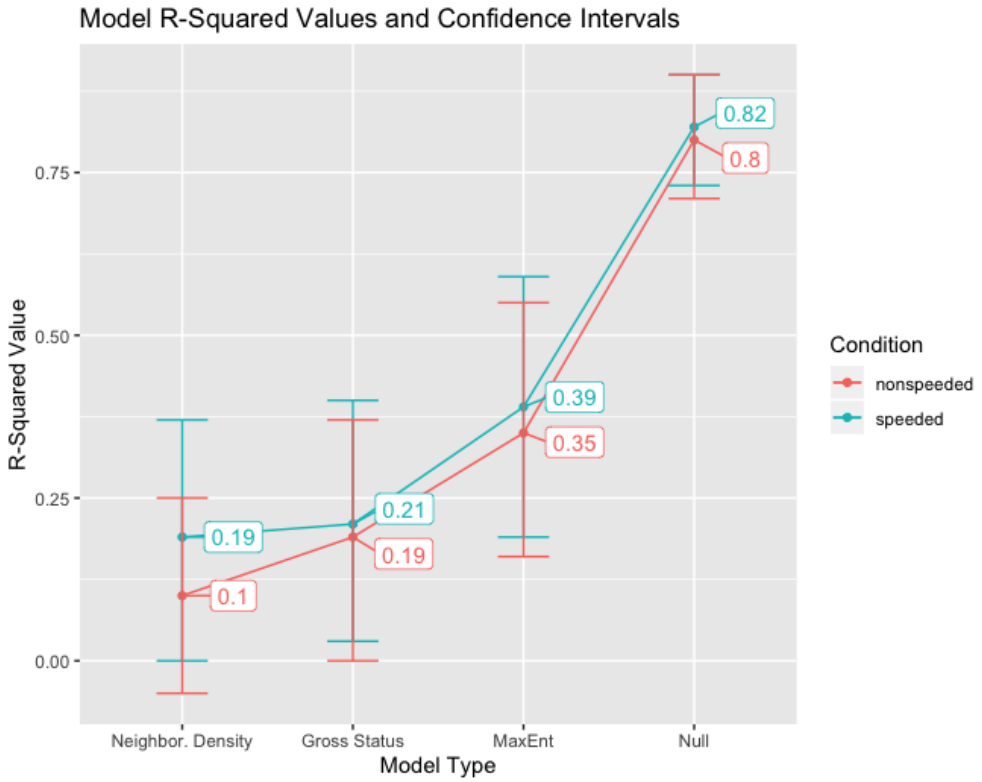

Figure 8: A plot showing the R-squared values for the speeded and non-speeded conditions in each of the different models, along with the confidence intervals for each R-squared value. This shows that all the R-squared values are captured within the confidence intervals for the opposing condition, meaning that there are no significant differences between conditions for any model.

In all cases, the R-squared values for both the speeded and non-speeded conditions are contained within the confidence intervals for both individual conditions, meaning that the change between the R-squared values for each condition can be viewed as non-significant. Also pictured is a null model, which only contains information about the means in the data. It is shown to have the best fit out of all of the models utilized here. Further details on this model will be discussed below.

Although it is not central my hypothesis, an investigation of which model fit best to the experimental data was also performed in order to compare the present study to previous ones. Table 2 below summarizes the R-squared values and confidence intervals for each of the models for only the non-speeded condition. This is because the non-speeded condition

is more comparable to previous studies than the speeded condition.

	Neighborhood Density	Gross Status	MaxEnt
R-squared Values	0.10	0.19	0.35
Confidence Intervals	-0.05 - 0.25	0.00 - 0.37	0.16 - 0.55

Table 2: R-sqaured values and confidence interval ranges for Neighborhood Density model, Gross Phonotactic model, and MaxEnt model of mean participant responses per stimulus **for the non-speeded condition only**.

The Neighborhood Density Model has the lowest R-squared values for both conditions and is therefore the worst-performing model to fit to this data. This means that lexical information does not contribute much explanatory power to the pattern of judgements here. The Gross Phonotactic Model does a bit better, showing that gross phonotactic information, which is a categorical phonotactic measure, improves our ability to explain the patterns in this judgement data compared to just using lexical information. The highest R-squared value obtained for both conditions is from the MaxEnt Model. Before checking levels of significance, this seems to indicate that gradient phonotactic information can explain more patterns in the judgements for this set of experiments compared to both of the other models. It is also clear that the speeded condition fits better to every model. Though it is important to point out that, according to Cohen's Rule of Thumb, all these effect sizes are quite small for behavioral experiment data (Cohen 1992). Furthermore, the confidence interval for the MaxEnt Model contains the value for the Gross Status Model (and vice versa) suggesting that the MaxEnt and Gross Status models are not significantly different from one another. The confidence interval for the Neighborhood Density Model also contains the R-squared value for the Gross Status Model (and again vice versa) meaning that the Neighborhood Density and Gross Status models are also not significantly different from each other. The only significant improvement is between the Neighborhood Density Model and the MaxEnt Model. Overall these comparisons should be interpreted with caution, since there are not very substantial differences between the fits for any of the models.

To examine this further, a null model which contains no predictors for either condition

was also performed. It produced an R-squared value of 0.82 for the speeded condition, and 0.80 for the non-speeded condition. This is a much larger effect size than any of the models mentioned above. This would imply that there are no systematic patterns in the data since a null model containing no predictors has a better fit to the data than any other model utilized in the analysis. However, this is strange since Figure 2 shows that the responses for both experimental conditions are highly correlated, suggesting that the responses from participants were not random and instead have some systematic behavior associated with them. This might imply that there is a model which has not yet been identified that would better explain the patterns that exist in this acceptability judgment data. This will be more thoroughly explored in the Discussion section.

From the investigation above, it can be concluded that differences between the model fits for the speeded and non-speeded conditions are not significant, and it was therefore not possible in my experiments to separate lexical knowledge from phonotactic knowledge using reaction time. Additionally, a null model produced the largest R-squared value in relation to this data, which along with the high correlation between the two experimental conditions, shows that there is a model that has not yet been entertained which may better explain the patterns in this data.

8 Post-Hoc Testing

In order to check how consistent responses were between individual participants, by-participant mean ratings for each stimulus is plotted below in Figure 9. Ratings for each stimulus seem to be more or less consistent, with the ratings for most stimuli having a unimodal distribution. However, there are items that have a bimodal distribution, meaning that participant responses to that item were not as consistent as those with unimodal distributions. This might be evidence that there were a few stimuli which induced perceptual illusions for some participants more than others. However, if there is an effect from these stimuli, it is likely that the effect is small since the majority of participants behave similarly

in terms of their ratings for the majority of the stimuli. The idea of perceptual illusions in the stimuli is further investigated in the Discussion section.

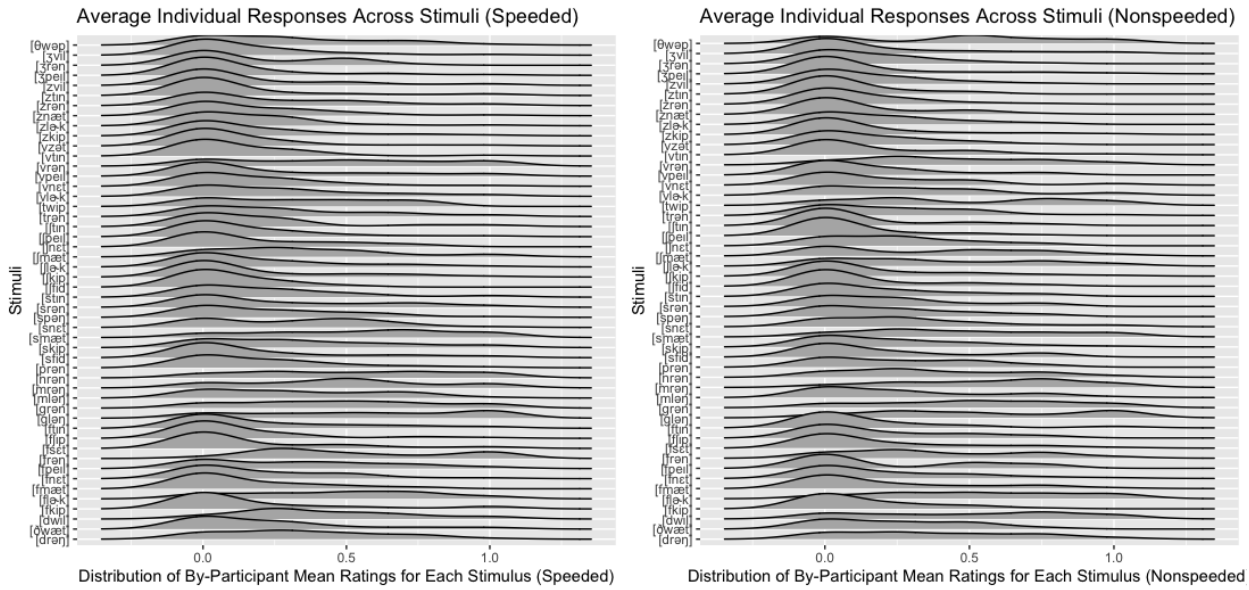

Figure 9: A ridge plot showing the mean ratings for each individual stimulus. These ratings are aggregated across all participants and plotted next to one another in order to discern how similarly individual participants behaved to one another between conditions for each stimulus item.

CHAPTER 5

DISCUSSION

Given that the difference in model fits between the speeded and non-speeded conditions are so small, I am not able to clearly confirm the hypothesis that I initially predicted. Although, there may be a few possible contributing factors which can be investigated in order to improve future iterations of this experiment. One potential issue could be that the behavior of online participants is inherently noisier than in laboratory conditions. The R-squared values for this data are smaller than those seen for in-person experiments. Without being able to supervise participants online, I am unable to verify if they are paying sufficient attention to the task, wearing proper headphones, running the experiment from a computer (not a mobile device) etc. It would be more ideal to repeat this experiment with participants in-person and under laboratory conditions post-COVID in order to replicate the results and verify the effects reported in section 4.

Related to rerunning the experiment, there are a few possible problems that could be improved upon in future versions of the experimental design that may reveal participant behavior which more closely resembles what was originally predicted for these tasks. One being that there is a design flaw in the speeded version of the experiment that did not give participants negative feedback when they were too slow to respond to the stimuli. Because of this, there are 695 total stimuli that do not have responses associated with them in the results of the speeded condition. By using feedback incorporated into the experimental design, I could enforce a quicker response time for all stimuli with less risk of losing out on responses in cases where participants lost track of how little time they had to answer.

There are also some potential improvements to be made with the recording and splicing of the stimuli. As outlined in section 3.1, I pronounced all the stimuli with a schwa vowel [ə] in between the two segments of the initial cluster. I then spliced it out in an attempt to remove any unintentional articulatory cues that may have been present if I had instead

attempted to simply pronounce the illicit cluster. However, in attempting to minimize those articulatory cues, it is possible that some sounds introduced in the stimuli were unnatural for English speakers. For example, in a cluster like [dr-], the [d] is actually pronounced more like an affricate than a stop. However, when I pronounced the segments [d] and [r] with the schwa between them, the [d] was pronounced just like a regular plosive [d]. This resulted in some clusters sounding different to how they would be heard in natural English speech. If participants picked up on the differences in these particular clusters, their ratings may have been influenced by the fact that the usually licit clusters sounded different than normal (and potentially less licit).

Additionally, some participants may not have heard the stimuli as intended. Since many of the stimuli contained illicit consonant clusters that are not usually present in English, it is possible that these clusters may have produced auditory illusions (Dupoux et al. (1999), Durvasula and Kahng (2016)). These illusions could have made the illicit clusters in the stimuli sound more like the English clusters which participants are more familiar with. This may have led to them rating some stimuli as 'good' because they were making judgments based on the illusory sequences, not the illicit clusters.

In order to check whether perceptual illusions were prevalent in this set of tasks, it is helpful to look at specific clusters that are prone to being misperceived by English speakers to see how the participants in this task responded to them. Davidson and Shaw (2012) identify 3 common perceptual illusions in English and the environments that they are most likely to occur. These include prothesis illusions (adding a sound to the beginning of o a word), most often found in fricative–initial sequences, deletion or change of the first consonant, often found in stop–nasal sequences, and vowel insertion, which is common in stop–stop sequences. There are not perfect examples of all of these environments in the stimuli used for this experiment, but close candidates can be observed to approximate these environments.

Many of the clusters occurring in these environments are not usually considered to be licit in English, so it would make sense for participants to rate them as 'bad' more often. If

it is instead the case that the stimuli containing these clusters are highly rated, there would be reason to believe that perceptual illusions interfered with participant's judgments in these tasks.

Fricative initial sequences like those in [fsɛt] and [zkip] have a neighborhood density of 4 and 1 (respectively), invalid gross status, and low ratings from participants in Scholes (1966). These factors would would predict that these stimuli be rated poorly by the participants in this study. Referring back to Figure 1, we can see that both stimuli are in fact rated quite low for both conditions. This tells us that prothesis illusions were probably not very prevalent for speakers in this set of tasks.

There are no stop-nasal sequences present in the stimuli for this experiment, but there are obstruent-nasal sequences, such as the fricative-nasal sequences in [vnɛt] and [fmæt]. This environment should trigger consonant deletion or change perceptual illusions. Both stimuli have low neighborhood density scores ([vnɛt] is 2 and [fmæt] is 4), they both have an invalid gross status, and were both rated poorly by participants in Scholes (1966). This again predicts that they should have received low ratings from participants in the present task. Ratings for [vn ɛt] are indeed quite low (around 20 total 'good' responses), however, the ratings for [fmæt] are almost double that of [vnɛt] at 41 total 'good' responses. The higher ratings for [fmæt] may point to consonant deletion or change illusions being more prevalent for participants in these tasks.

There are also no stop-stop clusters in the stimuli used in this experiments, but there are obstruent-obstruent clusters like those in [ftɪm] and [ʃp eɪl]. These kinds of clusters should trigger vowel insertion perceptual illusions. Both stimuli have a neighborhood density score of 2, an invalid gross status, and both are rated poorly by participants in Scholes (1966). This should again mean that these stimuli are rated low by participants in this set of tasks as well, and that turns out to be true for both conditions, meaning that illusory vowels do not seem to be a common perceptual illusion for this set of data.

To summarize this discussion of perceptual illusions, it is possible that consonant deletion

or change illusions were present for participants here, but no real evidence of vowel insertion or prothesis was found. Ultimately, it is difficult to verify whether these illusions are present or not without making changes to the experimental design. One way to address this would be to implement a stimulus transcription task into the experimental design after stimuli that have consonant clusters that are prone to being misperceived. Such a task would involve participants typing out exactly what they thought they heard after rating the the stimuli. This would allow us to verify whether participants are hearing the stimulus as it was presented, or it they are perceiving auditory illusions. Additionally, it could also function as way to check if participants are paying attention as they go through the experiment. Recognizing whether auditory illusions are present will help us to better understand the knowledge behind phonotactic judgements, since it will be clearer which sequences participants are actually using to make judgments. Having an accurate view of this may also change the way that the models fit to the data.

Considering the very small differences in the behavior of participants between the speeded and non-speeded trials, it is clear that it was not possible in my experiments to separate phonological and lexical knowledge using reaction time. However, the strong correlation between the two experimental conditions indicates that participant judgements are not random, meaning that there is another model that I have not yet explored which could better explain the patterns observed in this set of experiments. One possibility is the the Cohort Model (W. D. Marslen-Wilson and Welsh 1978), which predicts that the lexicon is activated at an extremely early stage in the process of speech recognition.[1] The Cohort Model predicts that the lexicon is almost immediately activated upon hearing an utterance, and all the possible words that the utterance can map to are activated at once. As the speaker hears more input, the possibilities for how many words that utterance might map to are narrowed. The main claim of the Cohort Model is that the recognition system is able to identify possible words in the lexicon so soon after the beginning of a word, that acoustic-phonetic input alone cannot

[1]Suggested by Louis Goldstein (personal communication).

be the only source of information in identifying sound sequences, even at early stages of perception.

The evidence for the above statement can be attributed to speech shadowing tasks used to measure the time it takes to recognize words in continuous speech contexts (W. Marslen-Wilson 1973; W. D. Marslen-Wilson 1975). The results of these tasks show that words could be accurately identified and responded to in 250-275 ms. Marslen-Wilson assigns about 75-100 ms of this response time to processes involved in response integration and execution, meaning that participants were initiating their responses between 150 to 200 ms after the beginning of each word. Recall that the mean of the reaction times in the speeded version of the experiment presented here is 934 ms, which is much later than the point that the lexicon is hypothesized to be activated by the Cohort Model.

According to this model, even though there is a large difference between the mean reaction times for the speeded and non-speeded conditions, the lexicon is activated too quickly in the process of speech perception for its effects to be eliminated at earlier reaction times. The Cohort Model might even lead us to make the opposite prediction from my initial one based on Fox's (1984) findings - that the more time participants have to make a decision about the utterance the have heard, the more possible candidates which the utterance could map to are ruled out. This would mean that participant behavior is *less gradient* in the non-speeded condition than in the speeded one.

Although the Cohort Model's predictions are based on real word recognition, there is no reason believe that the lexicon would not also be activated at the same speed in a nonce word context. If the activation of the lexicon is as integral to identifying possible words as the Cohort Model claims, it should also be active in a phonotactic acceptability judgement task which asks participants to evaluate how 'good' of an English-sounding word a nonce word is. In order to determine whether this is a more appropriate model for this set of tasks, a computational implementation of the Cohort Model would be necessary to compare with the models outlined in section 4.2.1, as well as the null model. Regardless, it is clear

that more research is necessary to understand the implications of this model for phonotactic acceptability judgments.

CHAPTER 6

CONCLUSION

In an attempt to attribute gradience in phonotactic judgments to lexical information rather than phonotactic information, I performed two variations of an acceptability judgment task. One pushed participants to provide their judgments as quickly as possible, and one allowed participants to provide their judgments at their own pace (inspired by Fox (1984)). I predicted that the level of gradience in the speeded condition would be lower due to a lessened availability of lexical information.

This appeared to be confirmed when observing the spread of the data for each condition. The speeded condition had a lower standard deviation (0.127) than the non-speeded condition (0.133). However, as discussed in section 4.2.1, this is not a completely objective measurement of gradience. There could be many factors that contribute to the level of gradience in participant responses that are not related to the source of their phonotactic knowledge. To avoid mistaking gradience due to unrelated factors as significant, mean participant responses for each stimulus were fitted to three different models which make various predictions about the source and level of gradience in the phonotactic grammar. The resulting R-squared values from those models showed that there were no significant differences in the level of variation between the speeded and non-speeded condition, indicating that lexical and phonotactic knowledge cannot be separated via reaction time. This would suggest that lexical access is still present at early reaction times, supporting claims about early Ganong Effects (Rysling et al. 2015; Kingston et al. 2016).

The effects sizes for each of the models was also examined and revealed that the MaxEnt Model (utilizes gradient phonotactic information), and the Gross Phonotactic Model (utilizes categorical phonotactic information) performed similarly, while the Neighborhood Density Model (utilizes lexical information), performed slightly worse. A null model with no predictors had the largest effect size out of all the models tested, suggesting a lack of

systematic patterns in the results. However, a strong correlation value for response patterns between conditions implies that the pattern of responses here is not random, but that there may be an additional model that would better explain the patterns in participant judgments. The Cohort Model does well in explaining the patterns observed in the present study, and it is possible that more research which considers this model in relation to phonotactic acceptability judgments could be fruitful. It also confirms that varying participant response time is not a productive means of separating the effects of phonotactic and lexical influence on phonotactic acceptability judgments.

Regardless of the results of this set of experiments, the significance of this study is that it explores a way to vary methodology in experimentation to contribute to theoretical views of phonology. Specifically, it attempts to simplify the theory of phonology by finding ways to show how certain effects (like gradience in judgments of phonotactics) may not need to be captured and explained by the phonological grammar. They can instead be explained by other sources, such as the lexicon, in order to lessen the scope of variation that the theory of phonology needs to account for.

APPENDIX

APPENDIX

#	IPA	CMUPD	Neighborhood Density	Gross Status	Rating
1	[gɹən]	G R AH1 N	18	valid	33
2	[stɪn]	S T IH1 N	18	valid	33
3	[smæt]	S M AE1 T	13	valid	32
4	[pɹən]	P R AH1 N	11	valid	32
5	[flɚk]	F L ER1 K	11	valid	31
6	[dɹəŋ]	D R AH1 NG	7	valid	31
7	[tɹən]	T R AH1 N	12	valid	31
8	[fɹən]	F R AH1 N	12	valid	29
9	[snɛt]	S N EH1 T	7	valid	28
10	[spəŋ]	S P AH1 NG	11	valid	28
11	[gləŋ]	G L AH1 NG	9	valid	27
12	[mɹəŋ]	M R AH1 NG	1	invalid	27
13	[ʃlɚk]	SH L ER1 K	4	invalid	22
14	[skip]	S K IY1 P	15	valid	20
15	[vɹən]	V R AH1 N	4	invalid	19
16	[sɹən]	S R AH1 N	9	invalid	14
17	[vlɚk]	V L ER1 K	2	invalid	14
18	[məŋ]	M L AH1 NG	4	invalid	13
19	[ʃtɪn]	SH T IH1 N	3	invalid	13
20	[fpeɪl]	F P EY1 L	4	invalid	13
21	[ʒɹən]	ZH R AH1 N	4	invalid	11
22	[fʃɪp]	F SH IH1 P	2	invalid	11
23	[ʃnɛt]	SH N EH1 T	2	valid	10
24	[ftɪn]	F T IH1 N	2	invalid	10
25	[zɹən]	Z R AH1 N	5	invalid	8
26	[nɹən]	N R AH1 N	5	invalid	8
27	[ʃmæt]	SH M AE1 T	1	valid	7
28	[sfɪd]	S F IY1 D	7	valid	7
29	[zlɚk]	Z L ER1 K	2	invalid	6
30	[ztɪn]	Z T IH1 N	1	invalid	6
31	[fsɛt]	F S EH1 T	4	invalid	6
32	[vzət]	V Z AH1 T	1	invalid	6
33	[ʃfɪd]	SH F IY1 D	1	invalid	5
34	[znæt]	Z N AE1 T	1	invalid	4
35	[fnɛt]	F N EH1 T	2	invalid	3
36	[fkip]	F K IY1 P	1	invalid	3
37	[vtɪn]	V T IH1 N	2	invalid	3
38	[zvip]	Z V IY1 L	2	invalid	3
39	[fmæt]	F M AE1 T	4	invalid	2
40	[ʃpeɪl]	SH P EY1 L	2	invalid	2
41	[vnɛt]	V N EH1 T	2	invalid	1
42	[ʃkip]	SH K IY1 P	2	invalid	1
43	[ʒpeɪl]	ZH P EY1 L	1	invalid	1
44	[zkip]	Z K IY1 P	1	invalid	0
45	[vpeɪl]	V P EY1 L	2	invalid	0
46	[ʒvɪl]	ZH V IY1 L	1	invalid	0
47	[dwɪl]	D W IH1 L	7	invalid	NA
48	[θwəp]	TH W AH P	0	invalid	NA
49	[twip]	T W IY1 P	5	valid	NA
50	[ðwæt]	DH W AE T	1	invalid	NA

Table 3: List of all stimuli used in this set of experiments along with their corresponding CMU Glyphs, neighborhood density scores, gross status, and previous rating provided by participants in Scholes (1966).

BIBLIOGRAPHY

BIBLIOGRAPHY

Albright, Adam (2009). "Feature-based generalization as a source of gradient acceptability." In: *Phonology*.

— (2007). "Natural classes are not enough: Biased generalization in novel onset clusters". In:

Albright, Adam and Bruce Hayes (2003). "Rules vs. analogy in English past tenses: A computational/experimental study." In: *Cognition* 90, pp. 119–161.

Armstrong, Sharon Lee, Lila R. Gleitman, and Henry Gleitman (1983). "What some concepts might not be". In: *Cognition* 13.3, pp. 263–308. ISSN: 0010-0277. DOI: https://doi.org/10.1016/0010-0277(83)90012-4. URL: http://www.sciencedirect.com/science/article/pii/0010027783900124.

Bailey, Todd M. and Ulrike Hahn (2001). "Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods?" In: *Journal of Memory and Language.* 44, pp. 568–591.

Berg, Hugo A. Van Den (2018). "Occam's Razor: From Ockham's via Moderna to Modern Data Science". In: *Science Progress* 101.3. PMID: 30025552, pp. 261–272. DOI: 10.3184/003685018X15295002645082.

Boersma, Paul and David Weenink (2016). *Praat: doing Phonetics by Computer [Computer program]*. Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/.

Chomsky, Noam and Morris Halle (1968). *The Sound Pattern of English*. New York, Evanston, and London: Harper and Row.

Davidson, Lisa and Jason A. Shaw (2012). "Sources of illusion in consonant cluster perception". In: *Journal of Phonetics* 40.2, pp. 234–248. ISSN: 0095-4470. DOI: http://dx.doi.org/10.1016/j.wocn.2011.11.005. URL: http://www.sciencedirect.com/science/article/pii/S0095447011001057.

Dupoux, Emmanuel et al. (1999). "Epenthetic vowels in Japanese: A perceptual illusion?" In: *Journal of Experimental Psychology: Human Perception and Performance* 25.6, pp. 1568–1578. ISSN: 1939-1277(ELECTRONIC);0096-1523(PRINT). DOI: 10.1037/0096-1523.25.6.1568.

Durvasula, Karthik and Jimin Kahng (2016). "The role of phrasal phonology in speech perception: What perceptual epenthesis shows us". In: *Journal of Phonetics* 54, pp. 15–34. ISSN: 0095-4470. DOI: http://dx.doi.org/10.1016/j.wocn.2015.08.002. URL: http://www.sciencedirect.com/science/article/pii/S0095447015000674.

Fox, Robert Allen (1984). ""Effect of Lexical Status on Phonetic Categorization"". In: *Journal of Experimental Psychology: Human Perception and Performance*.

Ganong, William F (1980). "Phonetic categorization in auditory word perception." In: *Journal of experimental psychology: Human perception and performance* 6.1, p. 110.

Gorman, Kyle (2013). "Generative Phonotactics". PhD thesis. University of Pennsylvania.

Hayes, Bruce and Colin Wilson (2008). "A Maximum Entropy Model of Phonotactics and Phonotactic Learning". In: *Linguistic Inquiry* 39.3, pp. 379–440. DOI: 10.1162/ling.2008.39.3.379.

Kingston, J. et al. (2016). "Eye movement evidence for an immediate Ganong effect." In: *Journal of experimental psychology. Human perception and performance* 42 12, pp. 1969–1988.

Marslen-Wilson, W. (1973). "Linguistic Structure and Speech Shadowing at Very Short Latencies". In: *Nature* 244, pp. 522–523.

Marslen-Wilson, William D. (1975). "Sentence Perception as an Interactive Parallel Process". In: *Science* 189.4198, pp. 226–228. ISSN: 0036-8075. DOI: 10.1126/science.189.4198.226. eprint: https://science.sciencemag.org/content/189/4198/226.full.pdf. URL: https://science.sciencemag.org/content/189/4198/226.

Marslen-Wilson, William D. and Alan Welsh (1978). "Processing Interaction and Lexical Access during Word Recognition in Continious Speech". In: *Cognitive Psychology* 10, pp. 29–63.

Palanab, Stefan and Christian Schitter (2018). "Prolific.ac—A subject pool for online experiments". In: *Journal of Behavioral and Experimental Finance*. DOI: https://doi.org/10.1016/j.jbef.2017.12.004.

Peirce, J. W. et al. (2019). "PsychoPy2: experiments in behavior made easy". In: *Behavior Research Methods*. DOI: 10.3758/s13428-018-01193-y.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rysling, Amanda et al. (2015). "Early Ganong effects". In: *ICPhS*.

Sarver, Isaac (2020). "A Systemic Evaluation of Computational Models of Phonotactics". MA. These. Michigan State University. DOI: https://doi.org/10.25335/5mch-cz06.

Scholes, R. J. (1966). *Phonotactic grammaticality.* The Hague: Mouton.

Shademan, Shabnam (2006). ""Is Phonotactic Knowledge Grammatical Knowledge?"". In: *Proceedings of the 25th West Coast Conference on Formal Linguistics.*

Vaden, K.I., H.R. Halpin, and G.S. Hickok (2009). *Irvine Phonotactic Online Dictionary, Version 2.0. [Data file].* Available from http://www.iphod.com.

Vitevitch, Michael S. and Paul A. Luce (1999). "Probabilistic Phonotactics and Neighborhood Density in Spoken Word Recognition". In: *Journal of Memory and Language* 40, pp. 374–408.

Weide, Robert L. (1994). *CMU Pronouncing Dictionary.* http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Wickham, Hadley et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.