

# UC Santa Cruz

## UC Santa Cruz Electronic Theses and Dissertations

### Title

How Unexpected: Exploring the Effect of Phonological Features on Perception of Sound Errors

### Permalink

<https://escholarship.org/uc/item/2t7933rw>

### Author

Miller Willahan, Claire

### Publication Date

2023

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nd/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

SANTA CRUZ

**HOW UNEXPECTED: EXPLORING THE EFFECT OF PHONOLOGICAL  
FEATURES ON PERCEPTION OF SOUND ERRORS**

A thesis submitted in partial satisfaction  
of the requirements for the degree of

MASTER OF ARTS

in

LINGUISTICS

by

**Claire Miller Willahan**

December 2023

The thesis of Claire Miller Willahan  
is approved:

---

Assistant Professor Amanda Rysling, Chair

---

Professor Jaye Padgett

---

Associate Professor Ryan Bennett

---

Peter Biehl  
Vice Provost and Dean of Graduate Studies

Copyright © by

Claire M. Miller

2023

# Table of Contents

<b>List of Figures</b>	<b>v</b>
<b>List of Tables</b>	<b>vi</b>
<b>Abstract</b>	<b>vii</b>
<b>Acknowledgments</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background</b>	<b>3</b>
<b>3 Experiment</b>	<b>11</b>
3.1 Methods .....	11
3.1.1 Materials .....	11
3.1.1.1 Item Design .....	14
3.1.2 Participants .....	17
3.1.3 Procedure .....	18
3.2 Results .....	20
3.2.1 Accuracy of transcription .....	21
3.2.2 Ratings .....	25
3.2.3 Decomposing ‘Manner’ .....	29
3.2.3.1 Accuracy of transcription .....	30
3.2.3.2 Ratings.....	32
3.3 Discussion .....	34

<b>4</b>	<b>Norming study: Completing the couplet</b>	<b>36</b>
4.1	Background .....	36
4.2	Methods .....	36
4.2.1	Materials .....	36
4.2.2	Participants .....	37
4.2.3	Procedure .....	37
4.3	Results .....	38
4.3.1	Accuracy by strength of association .....	39
4.3.2	Ratings by strength of association .....	40
4.4	Discussion .....	41
<b>5</b>	<b>General Discussion</b>	<b>43</b>
<b>6</b>	<b>Conclusion</b>	<b>47</b>
6.1	Future Directions .....	48
<b>A</b>	<b>Experimental stimuli</b>	<b>51</b>
<b>B</b>	<b>Manipulations by phoneme</b>	<b>58</b>
<b>C</b>	<b>Results from norming study</b>	<b>59</b>
C.1	Weakly associated binomials .....	59
C.2	Moderately associated binomials .....	61
C.3	Strongly associated binomials .....	65
<b>D</b>	<b>Additional transcription error tables</b>	<b>81</b>
	<b>References</b>	<b>82</b>

## List of Figures

3.1	Phoneme inventory of experimental sound manipulations, arranged vertically by place and horizontally by manner and voicing.....	14
3.2	Mean accuracy rates across mispronunciation conditions. Error bars show a 95% confidence interval.....	24
3.3	Ratings histograms for control and experimental conditions using the following rating scale: (1) <i>Completely unexpected</i> ; (2) <i>Mostly unexpected</i> ; (3) <i>Mostly as expected</i> ; (4) <i>Completely as expected</i> .....	26
3.4	Mean ratings across mispronunciation conditions. Error bars show a 95% confidence interval.....	28
3.5	Mean accuracy rates across manner features. Error bars show a 95% confidence interval.....	31
3.6	Mean accuracy rates across fricatives. Error bars show a 95% confidence interval.....	32
3.7	Mean ratings across manner features. Error bars show a 95% confidence interval.....	33
3.8	Mean ratings across fricatives. Error bars show a 95% confidence interval.....	34
4.1	Sample item from norming questionnaire [Above and beyond].....	37
4.2	Effect of strength of association and type of feature change on pronunciation judgments.....	40
C.1	Weakly associated binomials.....	59-60
C.2	Moderately associated binomials.....	61-64
C.3	Strongly associated binomials.....	65-80

## List of Tables

3.1	Sample binomial expressions.....	12
3.2	Imperfect contrasts.....	15
3.3	Experimental manipulations by target phoneme.....	17
3.4	Spectrum of errors.....	22
3.5	Accuracy of transcriptions by condition.....	23
4.1	Accuracy at different strengths of association.....	39
D.1	Transcription errors by condition.....	81
D.2	Manner transcription errors.....	81
D.3	Fricative transcription errors.....	81

## **Abstract**

How Unexpected:

Exploring the Effect of Phonological Features on Perception of Sound Errors

by

Claire Miller Willahan

The goal of this thesis is to better understand the impact of different phonological feature classes (voice, place of articulation, manner of articulation) on the likelihood of a listener recognizing a mispronounced word in natural speech. I examine this question by introducing mispronunciations into binomial expressions, semi-idiomatic phrases like *salt and pepper*, where the first half of the phrase lexically and semantically primes the second half. Mispronunciations were produced by deliberately changing the voice, place, or manner features of the onset consonant of the third word in a binomial. A set of experiments investigates (i) how listeners rate the effect of different feature errors on overall pronunciation quality, and (ii) how accurate they are at correctly recalling those erroneous pronunciations. Results are first analyzed for an effect of mispronunciation and find that listeners give higher ratings and are more accurate at correctly recalling binomial expressions with no mispronunciation, regardless of the particular feature change. A subsequent set of post-hoc analyses of the results are run, comparing the different mispronunciation conditions against each other and breaking down and comparing the distinctive



features that comprise the manner feature class (continuancy, stridency, nasality, lateralization). Results of the ratings and recall task find that different feature changes differently impact listener perception of sound errors: listeners are less likely to make mistakes when recalling words with manner errors in the target position than words with voice or place errors, and they assign lower ratings to binomial expressions with manner feature errors than equivalent binomials with mispronunciations involving voicing features.

## Acknowledgments

The truth of the matter is that it would take another, even longer thesis to detail every person to whom I owe a debt of gratitude for their support over the past few years. Since that is neither reasonable nor feasible, I will try to be as brief as possible.

Firstly, I would like to thank Amanda Rysling for providing an endless supply of inspiration, support, and patience since we first met. I could not have asked for a better advisor. Thank you for the years of open conversations and discussions about school and life that I know will guide me into and through the world beyond academia. I am so grateful that the timing worked out, and that it was you teaching Phono A that fateful Fall 2020 quarter a lifetime ago. Everything else I would like to say here would come across as overly gushy and emotional, so I will leave it at this: I am so grateful for the time I got to spend getting to know you and I will always be down to explore new restaurants, new foodstuffs, and the inner workings of the mind with you.

Thank you to my committee, Jaye Padgett and Ryan Bennett, for being extremely understanding as this thesis took much, much longer than I had originally intended. I am especially grateful that I had the opportunity to learn from both of you as your TA during my time in the UCSC Linguistics department. Thank you also to

Jaye and Grant McGuire for the constructive feedback that helped to reshape parts of this thesis.

Many thanks to Niko Webster and Taijing Xiao, my covid-cohort. Thank you for the hours we spent together-while-distanced in the first year of our program as we figured out how to feel less alone while wading through graduate school in the first years of a global pandemic. Thank you for all of the years that followed, as well.

Thanks also to the other members of our first-year classes and Discord: Mikkell, Eddie, Karen, and Sabrina. Exceptional thanks to Sabrina Madden, my best friend in the department; thank you for all those hours spent trying to crack the code in Syntax I and II, and the many conversations about life, the universe, and linguistics. Now that this is over, let's actually spend time together in Santa Cruz.

Special thanks to Jed Pizarro-Guevara for your kindness, friendship, and advice, and for being such a welcoming and supportive member of the department for the years that our time at UCSC overlapped (and even after you left). Your belief in your students (me, in this instance), is a large part of why I became so involved in the department and ended up in the BA/MA program. Thanks also to the group of Linguistics undergraduates, especially Gabby Pleasant, who welcomed me into their study group and made UCSC feel like a place I could still belong as a late-bloomer, readmit student. And many thanks to Gwyn Vandevere, who helped me to navigate an unreasonably complex readmission procedure and later gave me an opportunity to

pass on what I had learned as a peer advisor: your warmth and patience are a blessing to every student who walks through your door.

I would also like to thank every member of the faculty and graduate student body for their feedback, kindness, and guidance over the years. It has been a privilege to learn from every one of you.

Particular gratitude is owed to Max Kaplan for responding to about a thousand random questions at all hours of the day and night, including, but certainly not limited to, figuring out the solution that helped me finally get my experiment up and running. Thank you also for the many reassurances and non-linguistic conversations.

Thank you to Marzelle Addy, for your friendship always, and for feeding me and keeping me on track during the final months of this endeavor. Thank you especially for everything above and beyond in the final weeks of writing: the bullying wake-up calls, check-ins, passive-aggressive purchases on my behalf, and reminders to laugh and eat. To be loved is to be known.

Thanks also to Makenzie, Hilary, Carissa, Jenn, Erin, Frankie, Carly, Codi, and Maddie, for your unwavering friendship and support, and for keeping me sane throughout all of this. Special credit for that final list item goes out to the two group chats, *ghoulies* and *Van*, that supplied me with little packets of happiness in the form of messages and memes at all hours of the night.

Though they will never read this manuscript, I must also thank Jesse, Napeequa, and Rupert for the kind of love and emotional support only non-human friends can provide.

I would also like to thank Anna Paganelli, who helped me find the strength to come back to university at the age of 30, and to stick with it through the end of the world.

Thanks also to my mother and stepfather for their limitless support and love over the course of my life and, more particularly, while I was working on this project; I'm sorry that it would sometimes be months between phone calls. Endless thanks to my father for being a model of kindness, patience, and support, and for giving me invaluable insight into the world that is higher education. You are, and will always be, my hero.

Finally, thank you to Patrick for all of the things, big to small. In the infinity of possible utterances there is still not one adequate enough to express how grateful I am for all of the ways you have carried me over and through the past three years. I dedicate this thesis to you.

# 1 Introduction

Speech errors have long been used to make claims about speech production and speech perception, including but certainly not limited to the psychological reality of different linguistic concepts such as distinctive features. As previous research has found that listeners give different perceptual weight to different feature classes in word recognition (Martin & Pepperkamp, 2015), my goal in this paper is to examine the relative importance of different feature classes on the likelihood of recognizing sound errors. I designed a two-part experiment to find out which factors make a sound error more or less likely to disrupt a listener's understanding of the intended word to the point that they actually notice the error and are able to recall it.

There is a substantial body of work demonstrating that listeners are quite bad at accurately hearing and processing errors in natural speech (Alderete & Davies, 2019; Alderete & Tupper, 2018; Cutler, 1981; Ferber, 1991), to the point that even trained listeners detect about one in three errors in running speech (Ferber, 1991). This means that listeners are rapidly autocorrecting approximately two thirds of all the errors they hear. Based on this extremely high level of autocorrection, I wanted to test whether different feature changes made a sound error more or less likely to be detected, or if they were all equally likely to be missed.

Much of the previous research on speech errors has involved testing sounds in isolation (Martin & Pepperkamp, 2015; Frisch & Wright, 2002; Marin, Pouplier, & Harrington, 2010), or embedded in much larger stories or recordings of full conversations (Alderete & Tupper, 2018; Cole et al., 1978; Ferber, 1991). I wanted to

avoid longer recordings in the experiment in part to avoid a possible effect of duration leading to a decrease in accuracy. However, while it can be very useful to test single words outside of the context of running speech, it is unclear whether the results of tests involving such stimuli are ecologically valid when applied to speech as it is normally processed: language is usually situated in contexts that allow listeners to generate expectations about what a speaker is going to say, causing the listener to subconsciously autocorrect when they hear something wrong in the signal. In order to understand whether different feature changes are more robust and likely to be heard in natural speech, I needed a way to quickly generate strong expectations about the upcoming material.

Based on recent work by Delaney-Busch et al. (2019), which found that listeners are much faster at processing the meaning of a word when it is activated by a preceding, predictable context as small as a one-word semantic prime, I chose to use semi-idiomatic binomial expressions as the vehicle for my sound errors. Binomial expressions are multi-word expressions of the form “X and/or Y,” where the first half of the phrase lexically and semantically primes the second half. The most easily recognizable English binomial is probably *salt and pepper*, such that when a person has uttered the words “salt and...” the listener has already begun to anticipate “pepper.” Taking the fact that people tend to autocorrect two-thirds of all errors in natural speech with additional research showing that listeners are even more likely to incorrectly restore a word to its correct pronunciation when they have high expectations about what is coming next (Marslen-Wilson, 1975), it is possible that the

use of binomial expressions with such strong relationships between the component words will make it so listeners fail to hear the mispronunciations at all. If, however, the results find that people are differently sensitive to changes across different feature types, even in the location of a word about which they are able to build very strong expectations, this suggests that the different feature effects are more representative of general listening than simply testing words in isolation, which listeners can focus exclusively on.

## **2 Background**

The purpose of this thesis was to probe how changes to different phonological features affect listener perception of sound errors in predictable contexts where the increased expectation may cause listeners to autocorrect the mispronunciation and not notice the presence of an error at all. In these environments, is there a detectable difference between how disruptive a mispronunciation involving a change to the voice, place, or manner of articulation features is to the likelihood that a person will accurately hear the sound error?

To consider the effect of different mispronunciations on listener perception, I needed a way to create an expectation about what a word should sound like in order to measure the divergence between how a word “should” be pronounced and how it actually was pronounced. For this reason I created target stimuli by modifying English binomial expressions: three word phrases such as *salt and pepper* or *war and peace*, in which the last word of the expression is extremely predictable given the



preceding words in the phrase.<sup>1</sup> Binomial expressions are a type of formulaic language made up of two conjuncts that belong to the same part of speech (Benor & Levy, 2006) and are usually processed as whole, multiword expressions rather than as individual words (Eaton & Newman, 2018). Previous research has found that formulaic language is processed faster than comparable non-formulaic phrases in reading studies, and that the initial word in a binomial primes the final word (Carroll & Conklin, 2021). It is this priming effect of binomial expressions that makes them an excellent vehicle for testing listeners' predictions versus attention to actual input, because the priming of the second half of the binomial may make the intended final word so highly activated that listeners are even less likely to catch a feature change at all. I will use this to explore what makes a sound error more or less likely to be perceived. I modified the binomial expressions in this experiment by introducing 'faux-errors' into the onset consonant of the last word, since it has been observed that subjects are usually more successful at hearing sound errors when they affect the onset of a word (Cole et al., 1978). Each of these faux-errors, which I will simply refer to as 'errors' and 'mispronunciations' moving forward, differed from the canonical, expected production by minimal changes in the articulatory feature classes of place and manner of articulation and voicing.

The experimental stimuli in this study consisted of relatively strong, relatively frequent binomial expressions in order to test whether listeners catch feature-based mispronunciations even when they have generated very strong expectations about

---

<sup>1</sup> In this paper, predictability refers to the expectation for the final word in a binomial expression once the initial word has been heard.

what the pronunciation of the mispronounced word should have been. Initial frequency data were pulled from the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998), which collected word association, rhyme, and fragment norms. The USF Free Association Norms data were primarily used for the purposes of generating a sufficient number of binomial expressions for the experimental items and fillers used in the main experiment. However, because the data for these norms were collected from 1973-1998, the frequency and relationships between some of these words will have changed. To account for this possibility, I ran a separate study in which I showed subjects the first half of each binomial expression used as an experimental item in my main experiment, and asked them to complete it with the first word that came to mind (for further discussion, see Section 4).

By introducing an error into the onset of the final word of a binomial expression, this experiment hoped to exploit the tension between our bottom-up processing and top-down expectations. Our understanding of bottom-up processing is that listeners will pay the most attention to sounds in places that are most important to the signal, such as the beginning of a word (Cole et al., 1978), but, on the other hand, listeners' top-down expectations are strongest about highly predictable positions, like the last word in a binomial, and so those should be more prone to undergoing a process like automatic error correction where a listener subconsciously corrects an erroneous pronunciation based on what came earlier in the context of the sentence or phrase (Marslen-Wilson, 1975; Potter et al., 1993). The design of this study involved selecting binomial expressions whose conjuncts were strongly associated so the target

position should be one of high predictability, potentially triggering a much higher rate of automatic error correction than found in typical, non-formulaic speech.

Previous work by Martin & Pepperkamp (2015) examined the relative importance of different phonetic features (voice, place of articulation, and manner of articulation) in word recognition for French speakers via a mispronunciation detection task. They did this by asking their participants to press a key every time they recognized a correctly or incorrectly pronounced word. Martin & Pepperkamp's stimuli consisted of 35 correctly pronounced base items, each of which produced 4 mispronunciations by changing one feature in the obstruent onset of the base item.

Overall performance on the mispronunciation conditions was very low, evidence that participants had a difficult time recognizing words with even a one-feature change. There were, however, clear differences between performance on the types of feature changes, which Martin & Pepperkamp argue reflects the different degrees of importance that the different features have on word-recognition in French. Overall, Martin & Pepperkamp's experiment found that participants were much better at recognizing a stimulus with a voicing change than one with a change involving place or manner, thus voice changes had less of a negative effect on word recognition than place and manner changes. The authors argue that this is evidence for voicing features being less important for word recognition than place and manner features.

Many aspects of my experiment's design are derived from Martin & Pepperkamp's (2015) study on substitution errors in French. My experiment will also use similarly modified base items to study which feature changes affect listener

perception of sound errors in English, and whether any one of these changes is more disruptive to what people hear. If it is the case that the voice feature is also less important to American English speakers, I predict that mispronunciations involving voicing will be perceived less often than place and manner of articulation mispronunciations, and that subjects will regard them as less disruptive.

While not directly related to sound errors, Donca Steriade's research on the relative similarity between different feature contrasts makes predictions about which feature changes will be more or less disruptive to comprehension. For example, Steriade (2001) argued that a change in voicing is the most perceptually minimal possible change that could resolve violations of  $*[+VOICE/ \_ ]_{\text{Word}}$ , which explains why the languages of the world devoice word-final obstruents and don't, for example, change an underlyingly voiced word-final obstruent into a nasal or a glide.

Although Steriade (2001) focused on the relative perceptibility of different contrasts in word-final contexts, I wanted to see whether the same pattern of similarity judgments could be observed in different contexts, specifically in word-initial positions. It is generally understood that word-initial positions carry more perceptual information than their word-final counterparts and should, therefore, be more perceptually salient, making listeners less likely to miss or forgive an errorful feature change in onset rather than coda position. If this is the case then it is possible that Steriade's findings will not be applicable to the more salient context manipulated in my study. However, it is also possible that my results will show a change in voicing is still the most minimally perceptual feature change, even in onset positions.

Extending Steriade's arguments about the relative perceptibility of voicing changes in word-final positions to feature changes word-initially, one would predict that a voicing error will be less noticeable, and therefore less disruptive, than an error involving place or manner of articulation. These predictions align with not only Martin & Pepperkamp's (2015) results, but also with previous speech error research like Cole et al. (1978), who found that voicing errors in onsets tended to be caught less often during listening for mispronunciation tasks compared to manner and place errors.

Another potential explanation for why certain feature errors might be caught relatively more often has to do with how easily confusable those sounds are. Miller & Nicely's (1955) wideband noise-masking experiment involved discrimination tasks in which subjects heard a list of nonwords in the shape of simple consonant-vowel syllables (CVs) and were asked to respond with what C they thought they had heard at different signal-to-noise-ratios (SNRs). All responses were then coded up into what Miller & Nicely label confusion matrices: tables of how confusable different sounds in English are and what sounds they get confused for. Miller & Nicely were interested in understanding how noise and frequency distortion affected intelligibility of human speech. They analyzed perceptual confusions among 16 English consonants made up of plosives, nasals, and fricatives, which provided a system of five articulatory features that distinguished the different phonemes in the experiment. These five feature classes consisted of *voicing*, *place of articulation*, *nasality*, *affrication*, and *duration*. *Affrication* corresponds to the manner feature for continuancy, and *duration*

seems to correspond to stridency, because it is the feature that Miller & Nicely use to demarcate [s],[z], [ʃ],[ʒ] from other speech sounds. Taken with *nasality* these three represent all but one of the four features that constitute manner. Of these feature classes, voicing and nasality were the least affected by random masking noise than the others, and place of articulation was the most severely affected. Affrication and duration features were more robust than place, but less robust than voicing. Miller & Nicely's results lead to rather different predictions than Steriade or Martin & Pepperkamp: if the voice feature is the most robust and resistant to confusion, then the prediction would be that changes to the voicing feature will be more prominent and therefore regarded as worse, causing listeners to rate voicing changes lower than place and most of the manner features. Based on the same assumption, changes to this feature will stand out more and so listeners will be more accurate at correctly recalling mispronunciations in voicing. Miller & Nicely's results also predict that within the manner feature, changes involving nasal sounds will be the most prominent and will therefore be recalled more accurately and rated as worse than changes in stridency and continuancy.

In 2006, two researchers, Lovitt & Allen, ran an updated version of Miller & Nicely's 1955 experiment with very different results. In Miller & Nicely's original experiment, there were only 5 participants, a group of women from the Boston area who took turns being the talker and the listener. Instead of a homogenous group of talkers and listeners, Lovitt & Allen's version used stimuli from an online corpus and presented it to a group of people from around the University of Illinois

Champaign-Urbana. The thought behind this was that, unlike in Miller & Nicely, subjects would not already be familiar with the speaker's voice. Miller & Nicely mention having provided extensive training to their subjects, extensive enough that some people dropped out of the program and had to be replaced, which may have given subjects a chance to come up with systematic repair strategies for particularly confusing sounds (Lovitt & Allen, 2006, p. 2157). Lovitt & Allen provided their participants with some training during the experiment, but nothing extensive.

Although nasals still performed very well, the CVs with the most errors at the highest SNRs were those involving the voice feature, directly contradicting Miller & Nicely's results. Lovitt & Allen (2006) also proposed a new order within the confusion relationships, rejecting Miller & Nicely's order based purely on distinctive features. While each of the manner features (nasality, frication, and duration) were similar in both experiments, place and voice switched. These results may suggest predictions that are more in line with the rest of the literature: changes to place and manner features will be more prominent than voice changes. On the basis of Lovitt & Allen's results, we might also predict that nasality will be the most salient manner feature, like Miller & Nicely's original results.

The following section describes my main experiment, in which participants were asked to both rate and recall a series of correctly and incorrectly pronounced binomial expressions. Each of the mispronounced binomials was generated by a change in distinctive feature(s): place of articulation, manner of articulation, and voicing. A post-hoc analysis was then run in which the manner feature was separated

into its four independent features (continuancy, stridency, nasality, lateralization). These results should allow me to compare the representativeness of the two versions of the Miller & Nicely experiment, test whether Steriade's reasoning about contrasts in word-final positions can be extended to word-initial positions, and determine how different feature changes affect perception of sound errors and whether these effects correspond with their effect on word recognition.

## **3 Experiment**

### **3.1 Methods**

#### **3.1.1 Materials**

The materials for this experiment consisted of 60 binomial expressions from Standard American English (SAE), with each expression generating four possible pronunciations for a total of 240 experimental stimuli. 23 binomial expressions were taken from Morgan & Levy (2016), and an additional 37 were written for the purposes of this experiment.

I generated mispronunciations for each of the 60 control items by changing a single sound in the onset position of the second word in the binomial expression by as minimal a number of feature changes as possible (See Section 3.1.1.1 for a more thorough explanation), as allowed by the phoneme inventory of English. Each correctly pronounced control item thus yielded three mispronunciation conditions: one with a voicing feature change, one with a place feature change, and one with a manner feature change, for a total of 240 experimental items. The results of each of



these feature changes were treated as sound errors, with the base item the corresponding “correct” pronunciation. Whenever possible, the base phoneme was modified to generate a nonword of English (n = 201). When this was not possible, the resulting word was contextually inappropriate given the preceding word in the binomial (n = 39). Experimental stimuli were intentionally designed to obey English phonotactics as I wanted my “mispronounced” stimuli to reflect the shape of naturally occurring speech errors, which tend to be overwhelmingly phonotactically regular. (Alderete & Tupper, 2018).

The experimental items were distributed across four lists using a Latin Square so that 15 observations per condition were acquired per participant, with each participant seeing only one condition per item. A set of example items can be found in Table 3.1, and the full list of experimental items is provided in Appendix A. Each participant saw a total of 60 experimental binomial expressions randomly interspersed with 28 filler binomials, for a total of 88 binomials heard by each participant.

Table 3.1: Sample binomial expressions

<i>Control</i>	truth or dare	bread and butter	brother and sister
<i>Voicing error</i>	truth or tare	bread and putter	brother and zister
<i>Place error</i>	truth or gare	bread and tutter	brother and hister
<i>Manner error</i>	truth or zare	bread and mutter	brother and tister

The experiment included an additional 112 fillers, which were also binomial expressions of the form “X-CONJ-Y.” These fillers included 28 correctly pronounced

base items, 13 of which were modified versions of binomial expressions used in Morgan & Levy (2016), and 15 of which were generated by the experimenter for this experiment. Each base item was manipulated so as to generate three different mispronunciations. Like the experimental stimuli, the manipulation always affected the second half of the binomial expression but, unlike the experimental stimuli, the manipulation did not always involve the onset of the final word in the binomial, and the resulting nonwords did not always obey SAE phonotactics. Items with phonotactic violations were designed to be completely unexpected and confusing, and were included in an attempt to provide subjects with recordings that would be bad enough to fill out the lower end of the rating scale. To create these fillers I targeted various segments in the modified word to produce a mix of nonwords that disobeyed SAE phonotactics through the use of non-SAE phonemes (e.g. [ɲ] in ‘*sun and ñun*’ [sʌn ænd ɲun]), and phonotactically illicit onsets (e.g. the CC [ɹg] in *bride and rgoom* [bɹɑɪd ænd ɹgum]). Phonotactically regular fillers were generated by modifying the base word’s vowels and consonants across different syllable positions to contrast with the onset only manipulations in the experimental conditions. As with the experimental items, participants heard only one condition per filler binomial expression.

All base items, deliberate mispronunciations, and fillers were individually recorded by the author, a female speaker of California English, in Praat using a USB Blue Yeti microphone with an attached pop filter at a sampling rate of 44100 Hz. Stimuli were recorded at a speed as close to the natural speech rate of the speaker as possible to imitate how the phrases would be spoken in a normal conversation. The

average audio stimulus lasted 1113.57 ms. All recordings were scaled to an intensity of 55 dB.

### 3.1.1.1 Item Design

All test items were created through a deliberate mispronunciation of the final word in a binomial expression by changing its onset by as close to one feature as possible. The chart below (Figure 3.1) provides the place, manner, and voicing features for each of the consonants included in this experiment. Those consonants that only appeared in the mispronunciation conditions are highlighted in grey.

	Labial		Coronal			Velar		Laryngeal
	Bilabial	Labiodental	Dental	Alveolar	Post-Alveolar	Palatal	Velar	Glottal
Plosive	p b			t v			k g	
Nasal	m		θ ð	n				
Fricative		f v		s z	ʃ ʒ			h
Approximant				ɹ		j		
Lateral approximant				l				

Figure 3.1: Phoneme inventory of experimental sound manipulations, arranged vertically by place and horizontally by manner and voicing<sup>2</sup>.

Every target consonant (n = 14) in the experiment's inventory contrasts with another sound by one place feature change; eleven by one manner feature change, and ten by a voicing change. This leaves three sounds ([m], [n], and [ɹ]) without a

<sup>2</sup>Voiceless sounds are listed on the left side and voiced sounds on the right side of the cell, following the typical convention.

single-feature voicing contrast, and four without a single-feature manner contrast ([k], [g], [ð], [ʃ]). Due to the fact that this was a pilot study, and I wanted to collect data for as many contrasts as possible, I chose to keep those binomials whose final words began with one of the target consonants that did not allow a perfect, one feature change across the three experimental conditions. In order to fill out the missing conditions for each of these sounds, I appealed to other aspects of their articulation. A representative sample can be found in Table 3.2.

Table 3.2: Imperfect contrasts

<b>Condition</b>	<b>Place, voice</b>	<b>Place, manner</b>
<i>Control</i>	rise and ʃine	nature vs. nurture
<i>Voicing error</i>	rise and zine	<i>nature vs. turture</i>
<i>Place error</i>	rise and θine	nature vs. murture
<i>Manner error</i>	<i>rise and tine</i>	nature vs. zurture

Each of the consonants that lacked a voicing contrast involved a constriction at the same location in the oral cavity as another sound in the consonant inventory that did have a voicing contrast. To find voiceless counterparts to the two voiced nasal consonants I generalized from the fact that nasals are articulatorily similar to oral stops (also referred to more specifically as plosives): both involve a complete closure somewhere in the oral cavity, and their primary distinction lies in the fact that air may continue to flow through the nasal cavity in nasal consonants. The two nasal sounds found in this experiment, the bilabial [m] and alveolar [n], each have a corresponding voiced plosive in English, the voiced bilabial plosive [b] and the

voiced alveolar plosive [d]. As English does not have a voiceless alternative to [m] or [n], I replaced them with the voiceless bilabial plosive [p] and voiceless alveolar plosive [t], in their respective voicing mispronunciation conditions. I also contrasted the voiced alveolar approximant [ɹ] with the voiceless alveolar plosive [t] in its voicing mispronunciation condition for the reason that both involve a constriction at the alveolar ridge and neither are stridents.

Coming up with manner contrasts for sounds that only contrasted for place and voice was a little trickier. It was not possible to change the two velar stops (voiceless [k] and voiced [g]) into one or more different sounds by changing only their manner feature due to onset restrictions in SAE. Although [k] and [g] both contrast with the voiced velar nasal [ŋ], [ŋ] is not allowed as an onset by SAE phonotactics. Instead I chose to contrast both [k] and [g] with the voiceless glottal fricative [h], as they are all non-coronals produced via constrictions fairly far back in the oral cavity. The voiced dental fricative [ð] became the voiced alveolar nasal [n] as both are voiced coronal sounds, and the voiceless palato-alveolar fricative [ʃ] became a voiceless alveolar stop [t] because both are voiceless coronals.

While it is the case that not all phonemes of English may change into other phonemes via a change in just one feature, some phonemes have the option of changing their place and manner features in multiple ways. One such example is the voiced alveolar stop [d], which may become either a voiced bilabial stop [b] or voiced velar stop [g] through one change of its place feature. The voiced alveolar stop may

also become a voiced alveolar fricative [z], nasal [n], lateral-approximant [l], or approximant [ɹ] by simply changing its manner feature.

Although it was not always possible to generate multiple binomial expressions for every consonant, when there were multiple opportunities to change an item's place and manner features, the choice of which change to make was spread out over all possible phonemes. Preference was given for modifications that generated nonwords, but otherwise phonemes were selected so as to provide a well-rounded dataset. Table 3.3 provides a demonstrative sample of three of the most frequently occurring consonants across base items, and a full breakdown can be found in Appendix B.

Table 3.3: Experimental manipulations by target phoneme

<b>Control</b>	<b>Voicing</b>	<b>Place</b>	<b>Manner</b>
[d] = 10	[t] x 10	[g] x 6 [b] x 4	[l] x 3 [ɹ] x 4 [z] x 3
[s] = 8	[z] x 8	[f] x 3 [h] x 2 [θ] x 3	[t] x 8
[n] = 5	[t] x 5	[m] x 5	[d] x 1 [z] x 2 [l] x 1 [ɹ] x 1

### 3.1.2 Participants

Twenty-three individuals participated in this experiment. All were undergraduate students at UC Santa Cruz and received course credit for their participation. Each was randomly assigned to one of the four presentation lists. One participant did not fully complete the experiment and their data were excluded, leaving 22 participants whose data is considered here.

### **3.1.3 Procedure**

The experiment was conducted online and administered, unsupervised, on PCIBex Farm (Zehr & Schwarz, 2018). Participants were randomly assigned to one of the four lists, and each subject heard 60 target stimuli [i.e., one of the 60 base items presented in either the control condition (correct pronunciation), or in one of the three test conditions (voicing, manner, or place mispronunciation)] and 28 fillers in a randomized order. Subjects only ever heard one condition for each target and filler binomial expression.

The experiment began with an extensive training phase. In this phase, participants were guided through a series of exercises to introduce them to the concept of binomial expressions and the rating scale that they would be using in the experiment. Participants were told that they would be hearing common English phrases. While they were not directly taught about binomial expressions, subjects were told that the phrases they would hear in the experiment would follow a predictable format: two content words joined together by one of three conjunctions (and, or, versus) that were related by context and frequently co-occurred. Following the training phase, participants completed a brief practice round before moving on to the main task.

On each trial, participants heard one binomial expression and were prompted to rate how good the pronunciation was on the following four point scale:

1. Completely unexpected
2. Mostly unexpected
3. Mostly as expected

#### 4. Completely as expected

Since there are many ways that a pronunciation might be regarded as good or bad, participants were told to think of goodness and badness as a function of how disruptive the pronunciation was to their understanding of the phrase overall. If something was pronounced in an unexpected way, they were asked to consider how unnatural or unpredictable the pronunciation was compared to what they had expected to hear. A rating of *completely unexpected* was defined as a pretty bad pronunciation that was either very hard to understand or pronounced in an unexpected way. A rating of *mostly unexpected*, but not completely unexpected, should reflect a pronunciation that was poor but not awful; one or more of the words in the phrase was pronounced pretty badly, but the mispronunciation wasn't completely unnatural or confusing to the point that it became difficult to understand the words or what the phrase was supposed to be. *Mostly expected*, but not completely as expected, should be used to rate a phrase where one or more of the words was pronounced incorrectly but the mispronunciation wasn't confusing, or didn't affect the understanding of the phrase. Finally, participants were told to provide a rating of *completely as expected* when what they heard was obviously good, and when every word in the phrase was easy to understand and pronounced in a predictable way.

Participants were instructed to judge the quality of the pronunciation by how natural or predictable each of the words in the phrase were, given the context of the phrase as whole, such that if they heard a correctly pronounced word of English that



didn't make sense with the rest of the phrase, for example the phrase *hugs and hisses* in place of *hugs and kisses*, they should treat it as a mispronunciation.

Subjects saw a blank screen as the stimulus played, and were then asked to click a button rating how good the pronunciation of the phrase they just heard was. Subjects were allowed to proceed at their own pace, without limits on their response times. The experiment would only move forward after the participant provided a rating response via button press for the binomial expression that they had just heard.

Once the subject rated the binomial by clicking one of the four available buttons, the experiment immediately moved on to the next screen where they were asked to fill in a text box with their recollection of the phrase they had just listened to. Subjects were explicitly instructed to type out the words and non-words exactly as they had heard them, including any and all mispronunciations, entering their best guess if they were unsure. Subjects were again allowed to proceed at their own pace, without limits on their response times, and the experiment would not proceed until they pressed "Enter" or "Return." The audio for each subsequent trial would begin to play after a one second pause. Once subjects had heard all 88 binomials from their list they were asked to respond to a series of short answer questions about the task before returning to the experimenter's Zoom room for a debriefing.

## **3.2 Results**

Every subject performed at greater than 80% accuracy on control items, which meant no subjects were excluded from the analysis based on poor performance on

controls. One subject was excluded because they failed to correctly transcribe more than 50% of their stimuli across conditions, leaving data from a total of 21 subjects for analysis. Data from two control items (n.obs. = 11) were removed due to lower than expected ratings in over 50% of trials across participants. Additionally, data from four individual trials were removed because the subject reported that they did not hear the audio for those trials. Taken together, this resulted in a loss of 1.19% of observations for both the rating and the transcription tasks.

### **3.2.1 Accuracy of transcription**

Participants were overall very good at accurately transcribing the binomial expressions they heard, regardless of whether the target sound was correctly or incorrectly pronounced. Out of the 1245 responses that were collected, only fifteen percent (n = 181) were recalled incorrectly.

Responses were coded as incorrect when the target sound was not recalled correctly, whether by autocorrecting to the expected sound (e.g. responding with *cats and dogs* instead of *cats and togs*), hearing a sound that was not primed by the binomial (*action versus leality* rather than *action versus jeality*), or missing the unexpected sound completely (*analog or igital* instead of *analog or gigital*). Responses were also coded as incorrect when the error was so disruptive that it caused a participant to misunderstand the word or binomial entirely (examples include *orans teas* in place of *war and teace* and *pot of gold* instead of *hot or gold*). I included the latter in the list of mistranscriptions, even when the subject correctly reported the target sound, based on the following reasoning: if a listener was unable to

understand the binomial as a whole, then they would be unable to make predictions about the sounds they should expect in the latter half of the expression, and there could no longer be any tension between the sound they heard and the sound they expected. In multisyllable words, we normally find autocorrection in the second and third syllables, so if binomials are truly acting as one unit, it is reasonable to predict autocorrection in the third and final word of a binomial expression (Marslen-Wilson, 1975). Table 3.4 provides a breakdown of the different kinds of transcription errors found in the data.

Table 3.4: Spectrum of errors

Error type	Count
Autocorrected to expected sound	80
Changed binomial into another phrase	33
Completely missed phrase: nonwords	32
Heard unexpected sound	32
Deleted target sound	4
Total	181

The overall accuracy across conditions is reported in Table 3.5. Participants were overwhelmingly successful at correctly transcribing the control items. They still performed very well, but were notably less successful at correctly transcribing binomial expressions with errors in the target position. Manner feature errors were less detrimental to accurately reporting mispronunciations than errors in either voice

or place features and, with less than a percentage point's difference in accuracy, subject performance on items with voicing and place errors was essentially equivalent.

Table 3.5: Accuracy of transcription by condition

	<i>No error</i>	<i>Voice error</i>	<i>Place error</i>	<i>Manner error</i>
<b>Incorrect</b>	3	65	67	46
<b>Correct</b>	299	250	246	269
Accuracy	99.01%	79.37%	78.59%	85.39%

Accuracy data were fit to a generalized linear mixed effects model using the lme4 package (Bates et al., 2015) in R (R Core Team, 2023) with CONDITION (no error, voice error, manner error, and place error) as fixed effect and random effects for SUBJECT and ITEM. All three mispronunciation conditions were significantly different from the control condition: voice error [estimate ( $\beta$ ) = -3.6169, standard error (SE) = 0.5947,  $z = -6.082$ ,  $p < 0.001$ ]; place error [ $\beta = -3.7105$ , SE = 0.5954,  $z = -6.232$ ,  $p < 0.001$ ]; manner error [ $\beta = -3.0804$ , SE = 0.5977,  $z = -5.154$ ,  $p < 0.001$ ]. There was a main effect of CONDITION, with subjects demonstrating lower accuracy in each of the mispronunciation conditions relative to the control condition.

A post-hoc analysis was done comparing the relative impact of the different mispronunciation conditions on how well listeners perform at correctly hearing and recalling sound errors. Figure 3.2 shows the mean accuracy rates for each of the mispronunciation conditions. This measure allows us to compare the relative effect

that each of these feature changes has on how well participants did at correctly perceiving sound errors. There is an apparent difference between errors involving manner feature changes and errors involving changes in place or voice features, with participants being more accurate at hearing and recalling manner errors compared to the other two.

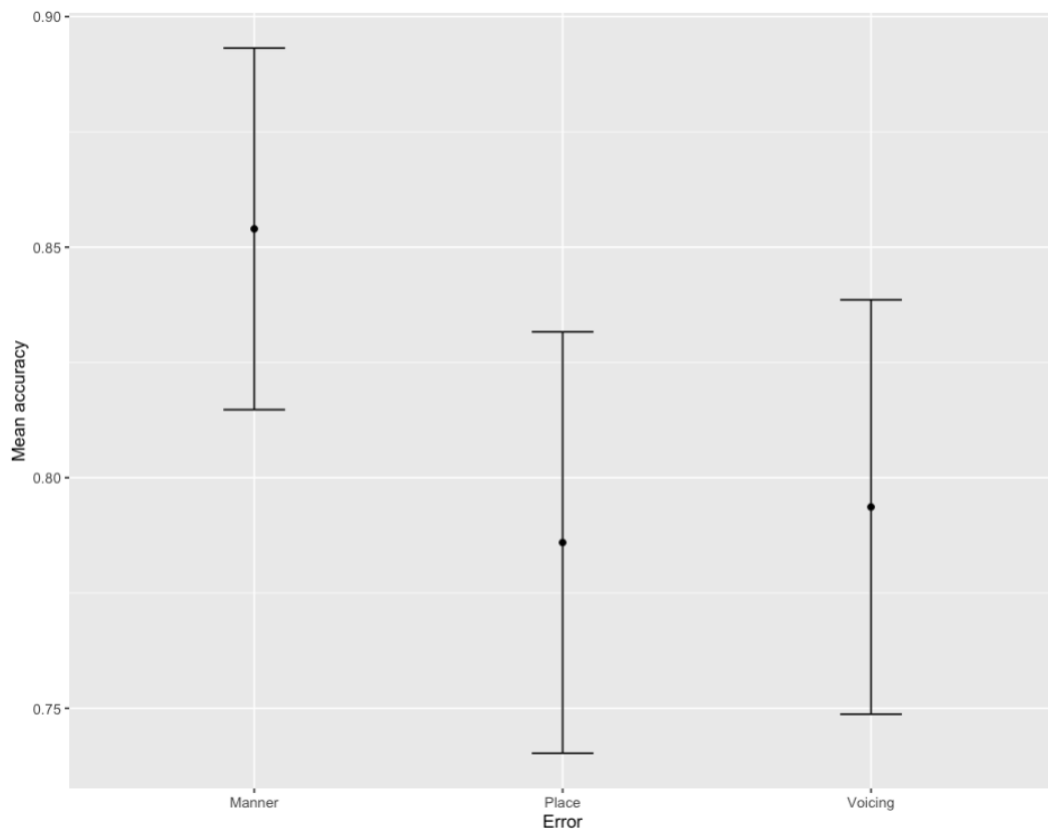


Figure 3.2: Mean accuracy rates across mispronunciation conditions. Error bars show a 95% confidence interval.

Generalized linear mixed effects models were fit to the accuracy data with EXPERIMENTAL CONDITION (manner error, place error, voice error) as a fixed effect and SUBJECT and ITEM as random effects. As might be expected, given the distribution of the confidence intervals in Figure 3.2, the manner mispronunciation

condition differed significantly from both the voice ( $\beta = -0.6301$ ,  $SE = 0.2286$ ,  $z = -2.756$ ,  $p < 0.01$ ) and place ( $\beta = -0.5364$ ,  $SE = 0.2281$ ,  $z = -2.352$ ,  $p < 0.05$ ) mispronunciation conditions, and they did not differ significantly from each other ( $z = 0.441$ ,  $p > 0.1$ ).

### 3.2.2 Ratings

Participants were asked to rate the different phrases they heard on the following four-point scale: (1) *Completely unexpected*, (2) *Mostly unexpected*, (3) *Mostly as expected*, (4) *Completely as expected*. The central tendency measures for participant ratings across different conditions mirrored the general pattern of the accuracy data, with participants assigning a median<sup>3</sup> rating of 4 to the control condition, 3 to both the voice and place mispronunciation conditions, and 2 to the manner mispronunciation condition. Impressionistically, these numbers show the different conditions having a similar impact on ratings as they did to accuracy, with errors in manner standing out more than errors involving place or voice features, place and voice errors having fairly equivalent effects, and controls standing out from the experimental conditions. This final impression is further supported by the fact that the mode value for the control condition was 4, and the mode value for all other conditions was 3. The ratings for each condition are shown in Figure 3.3. This histogram shows a large drop in the number of 4 ratings from correctly pronounced

---

<sup>3</sup>Because these ratings data are on an ordinal scale, and the distance between the different categories is therefore not consistent, I have chosen to only include median and mode values and not the mean or standard deviation. In an ordinal scale, variables have a natural and meaningful order, but the intervals between them are not numerically equal. Since both the mean and standard deviation assume equal intervals, applying them to an ordinal scale has the potential to misrepresent the data.

binomial expressions to mispronounced binomials, and then a steady decrease in 4 ratings from voice to place to manner. While the majority of the mispronounced items (regardless of feature class) were given a rating of 3, the histogram also shows the number of *2s* and *1s* steadily increasing from voice to place to manner.

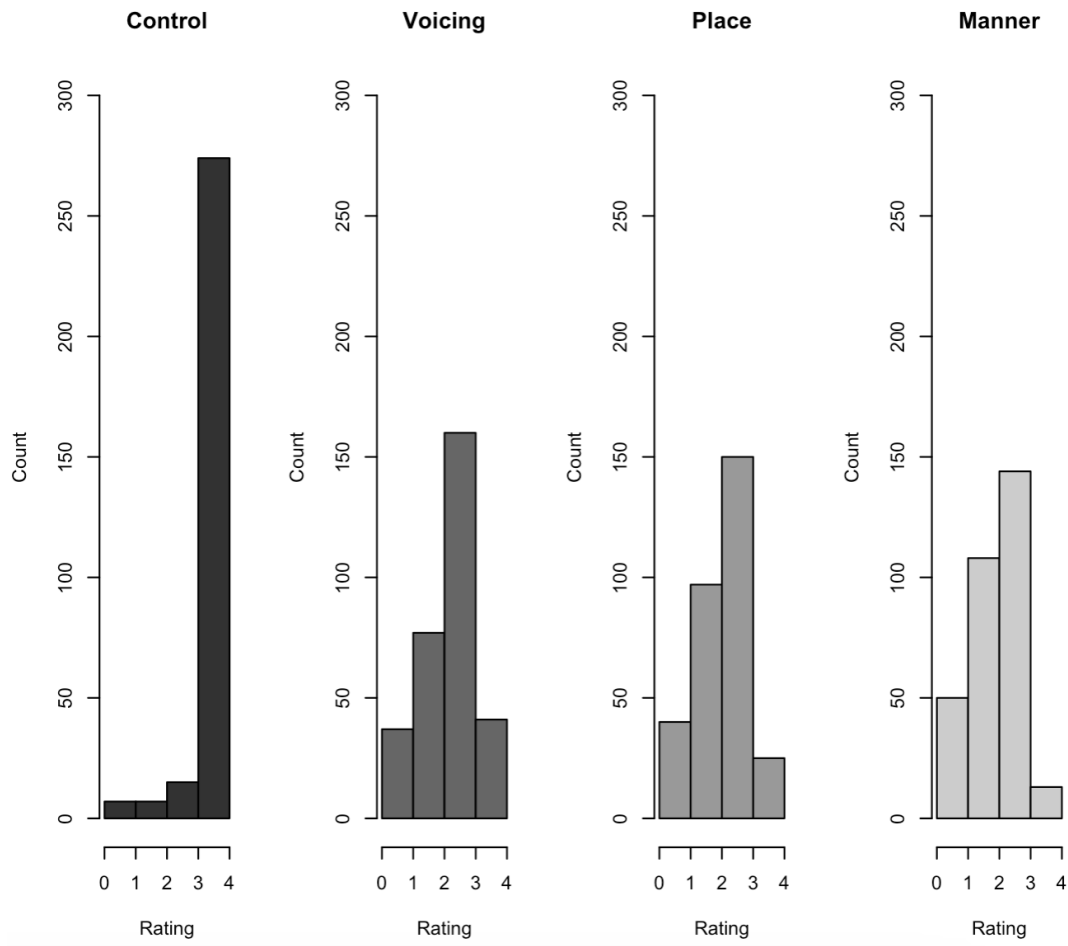


Figure 3.3: Ratings histograms for control and experimental conditions using the following rating scale: (1) *Completely unexpected*; (2) *Mostly unexpected*; (3) *Mostly as expected*; (4) *Completely as expected*.

A cumulative link mixed effects model was fit to the ratings data using the *Ordinal* package (Christensen, 2022) in R (R Core Team, 2023). Ratings were treated

as the dependent variable, CONDITION as fixed effect, and SUBJECT and ITEM as random effects. All three mispronunciation conditions were significantly different from the control condition: voice error [estimate ( $\beta$ ) = -8.418, SE = 1.590,  $z = -5.293$ ,  $p < 0.001$ ]; place error [ $\beta = -8.905$ , SE = 1.623,  $z = -5.486$ ,  $p < 0.001$ ]; manner error [ $\beta = -9.266$ , SE = 1.660,  $z = -5.583$ ,  $p < 0.001$ ]. There was a main effect of CONDITION, with subjects rating each of the mispronunciation conditions lower relative to the control condition.

Because I was also interested in comparing the effects of the different mispronunciation conditions against each other, a post-hoc analysis was run to compare the relative impact they had on listener ratings. Figure 3.4 provides 95% confidence intervals for the mean ratings of each mispronunciation condition. The confidence intervals for voice and manner errors do not overlap, with participants rating manner errors worse on average than voicing errors. Although there is some overlap between the confidence intervals for place errors and the other two, the data show ratings decreasing at a fairly consistent rate with voicing errors being regarded as less disruptive than errors driven by a change in the place feature, and place errors being rated as less disruptive than manner errors.



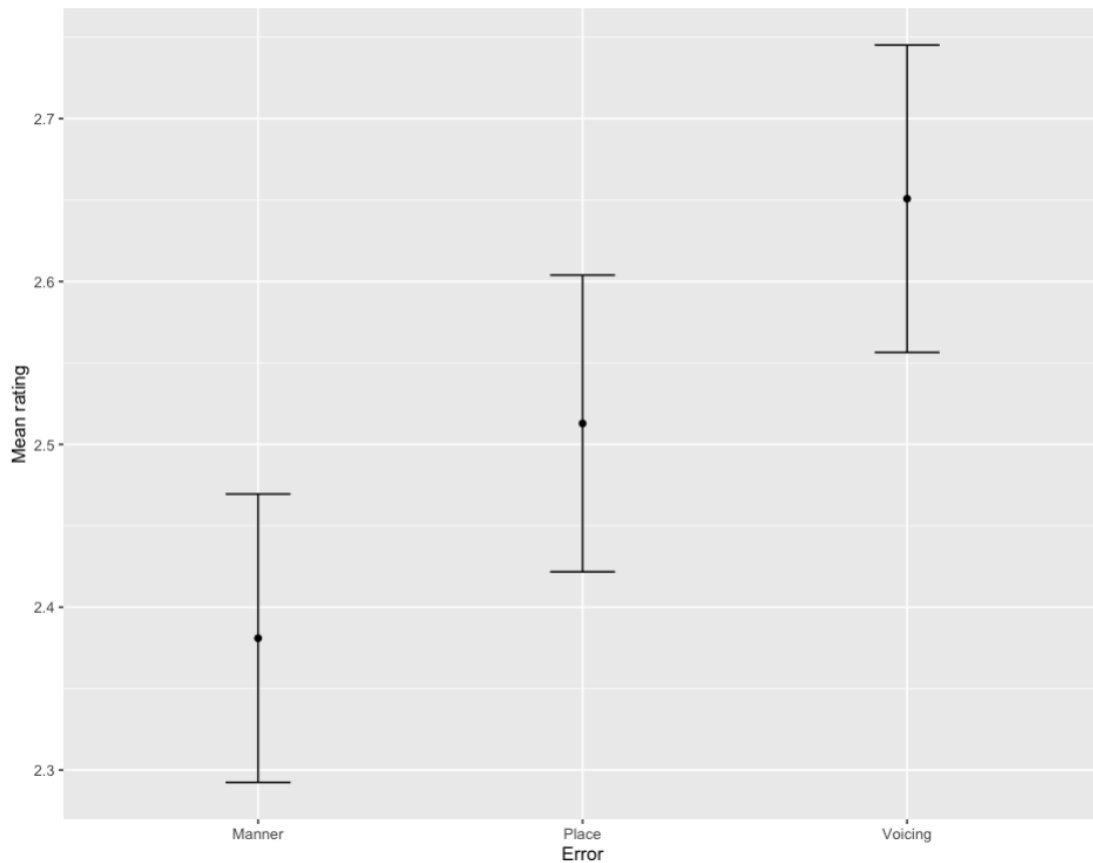


Figure 3.4: Mean ratings across mispronunciation conditions. Error bars show a 95% confidence interval.

A post-hoc analysis was done comparing each of the mispronunciation conditions against each other. A series of cumulative link mixed effects models were fit to the ratings data with each of the mispronunciation conditions treated as the baseline, EXPERIMENTAL CONDITION (manner error, place error, voice error) as fixed effect and SUBJECT and ITEM as random effects. The only model that yielded a notable result was the one that treated manner errors as the referent. In this model, ratings for manner differed significantly only from the voice condition ( $\beta = 0.8475$ ,  $SE = 0.2783$ ,  $z = 3.046$ ,  $p < 0.01$ ), with subjects rating manner errors as worse than voice errors. Although the confidence intervals in Figure 3.4 show a trend of subjects

rating manner of articulation errors as more unexpected than errors in place of articulation, this trend did not reach significance in the model (place:  $\beta = 0.3608$ , SE = 0.2308,  $z = 1.564$ ,  $p > 0.1$ ). Models fit to the ratings data with place of articulation as the baseline did not find a significant effect of EXPERIMENTAL CONDITION: ratings for errors in place of articulation did not differ significantly from those involving changes in voicing ( $\beta = 0.4867$ , SE = 0.2711,  $z = 1.795$ ,  $p > 0.05$ ) or manner ( $\beta = -0.3608$ , SE = 0.2308,  $z = -1.564$ ,  $p > 0.1$ ).

### **3.2.3          Decomposing ‘Manner’**

This section responds to a question that came up during a meeting with one of my committee members. When discussing the different articulatory feature classes used to more neatly divide up distinctions between sounds in the world’s phoneme inventories, it is easy to forget the fact that the distinctive features that make up the larger ‘manner’ class are much less homogenous than their place and voice counterparts. This is due to the fact that, while voicing and place differences are fairly well defined, the manner class is really an umbrella for four different feature classes. Because of this, ‘manner’ may stand for different things to different researchers. To Martin & Pepperkamp, whose experimental manipulations involved obstruents, it stood for changes in continuancy, but it can also stand for stridency, nasality, or lateralization. Because my experiment included more sound categories, manner was able to stand for continuancy, stridency, nasality and, to a smaller extent,

lateralization.<sup>4</sup> This section takes a look at the trends in the accuracy and ratings results across the different features that make up manner.

### 3.2.3.1 Accuracy of transcription

Figure 3.5 provides 95% confidence intervals for the mean likelihood of a correct response for each of the distinctive manner features. Please note that *lateral* has such a wide confidence interval because there were only four items (for a total of 16 observations) that involved a change in the lateral feature. There is a fair amount of overlap in the three remaining features, with people performing slightly better when the manner change involved a nasal feature, and slightly worse when stridency changes were involved.

---

<sup>4</sup>The lateral feature is relatively absent from my dataset because voiced alveolar lateral approximants were only able to generate mispronunciations for one of my experimental conditions due to facts about the phoneme inventory of SAE.

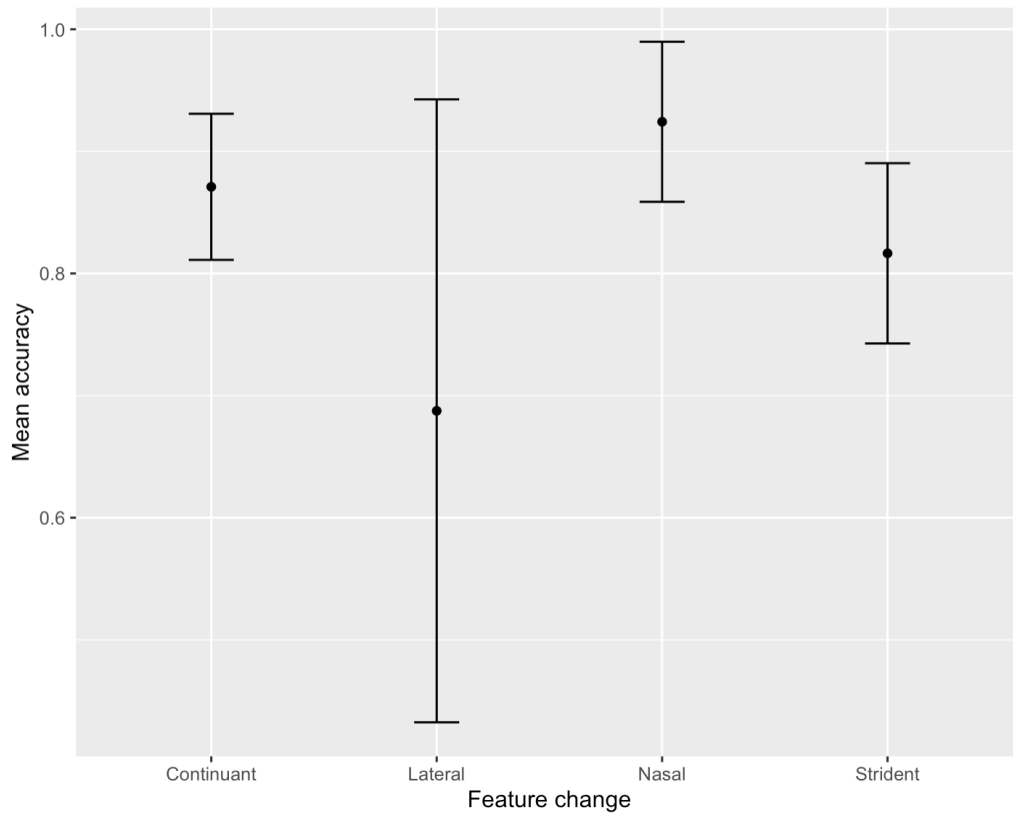


Figure 3.5: Mean accuracy rates across manner features. Error bars show a 95% confidence interval.

Figure 3.6 provides another set of confidence intervals for the mean likelihood of a correct response between mispronunciations involving a non-strident fricative (e.g. [f], [v]) and those involving a strident fricative (e.g. [s], [z]). There is a high rate of overlap, but the results show participants performing slightly better on items with mispronunciations involving non-strident fricatives.

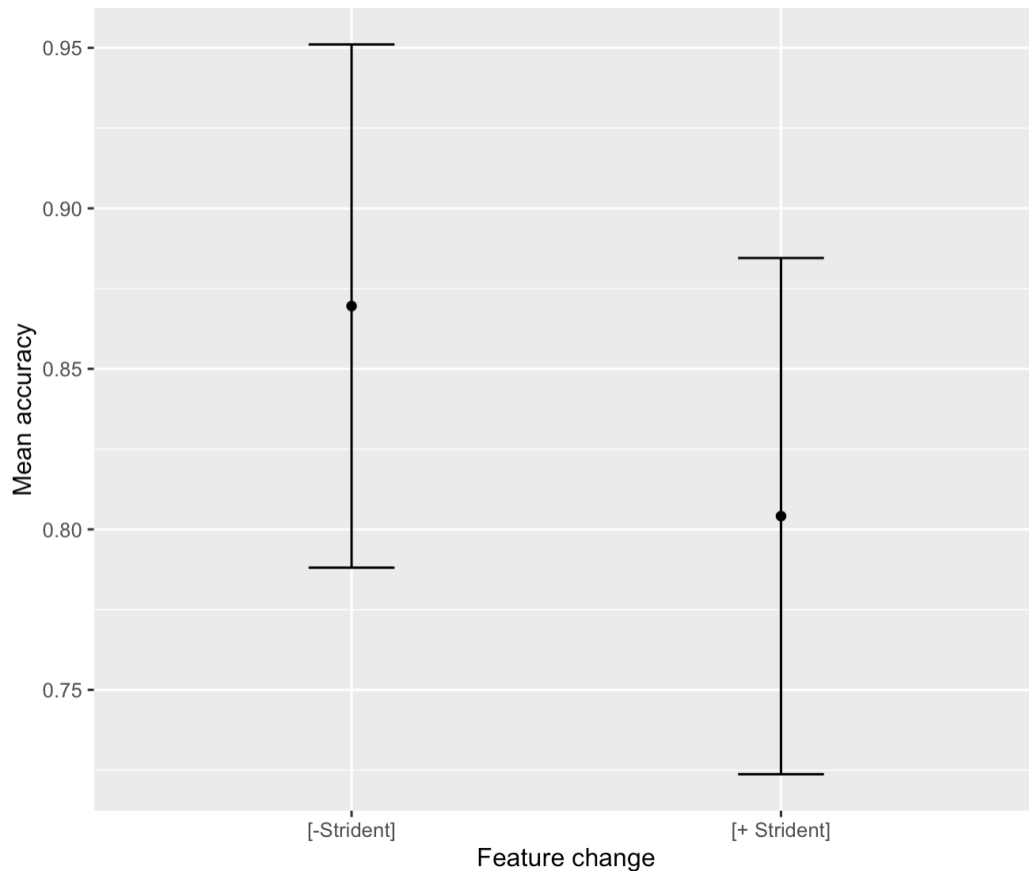


Figure 3.6: Mean accuracy rates across fricatives. Error bars show a 95% confidence interval.

### 3.2.3.2 Ratings

Figure 3.7 provides 95% confidence intervals for the mean ratings for each distinctive manner feature. Again, please note that the confidence interval for the lateral condition has such a wide confidence interval because of its very limited set. There is, once again, a fair amount of overlap between the continuant and nasal conditions with people rating the pronunciation as slightly better when the manner change involved a nasal feature. The stridency condition demonstrates less overlap in

this task, with participants rating mispronunciations of the strident feature lower compared to continuant and nasal mispronunciations.

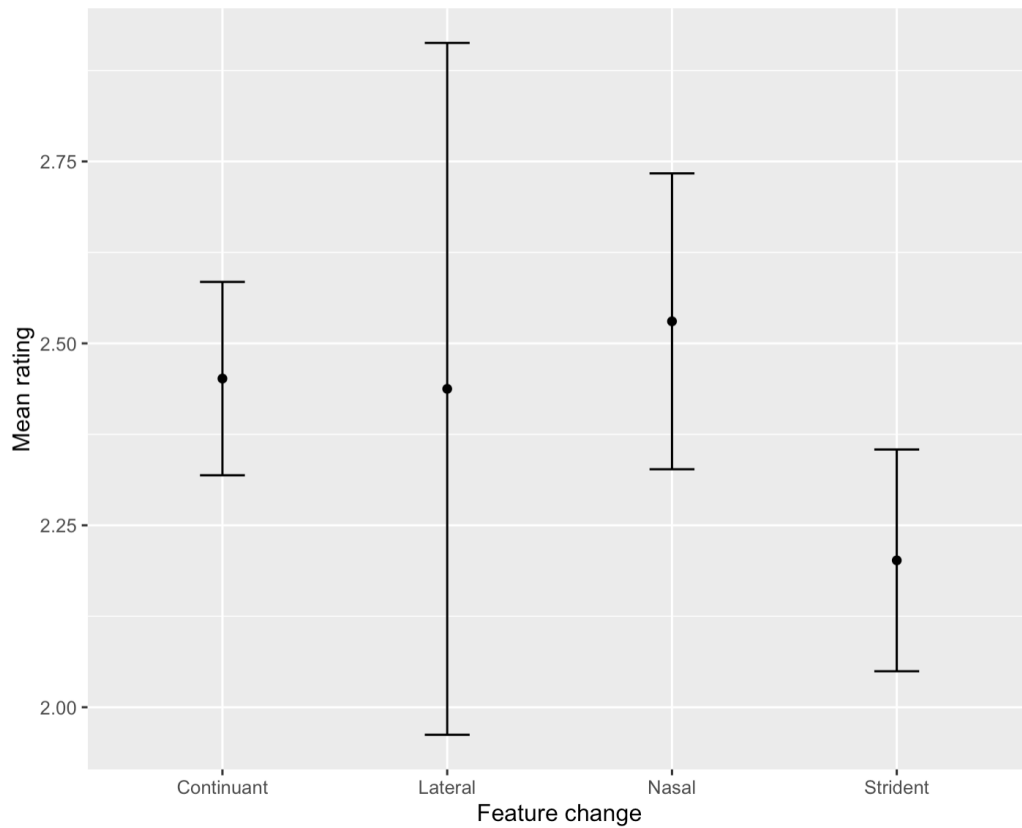


Figure 3.7: Mean ratings across manner features. Error bars show a 95% confidence interval.

Figure 3.8 gives the 95% confidence intervals for the mean ratings of mispronunciations involving non-strident fricatives and strident fricatives. There is some overlap, but the results show participants performing better on items with mispronunciations involving non-strident fricatives. There is less overlap between the two intervals when compared against the accuracy data, with participants rating items with strident fricative mispronunciations as worse than items with non-strident fricative mispronunciations.

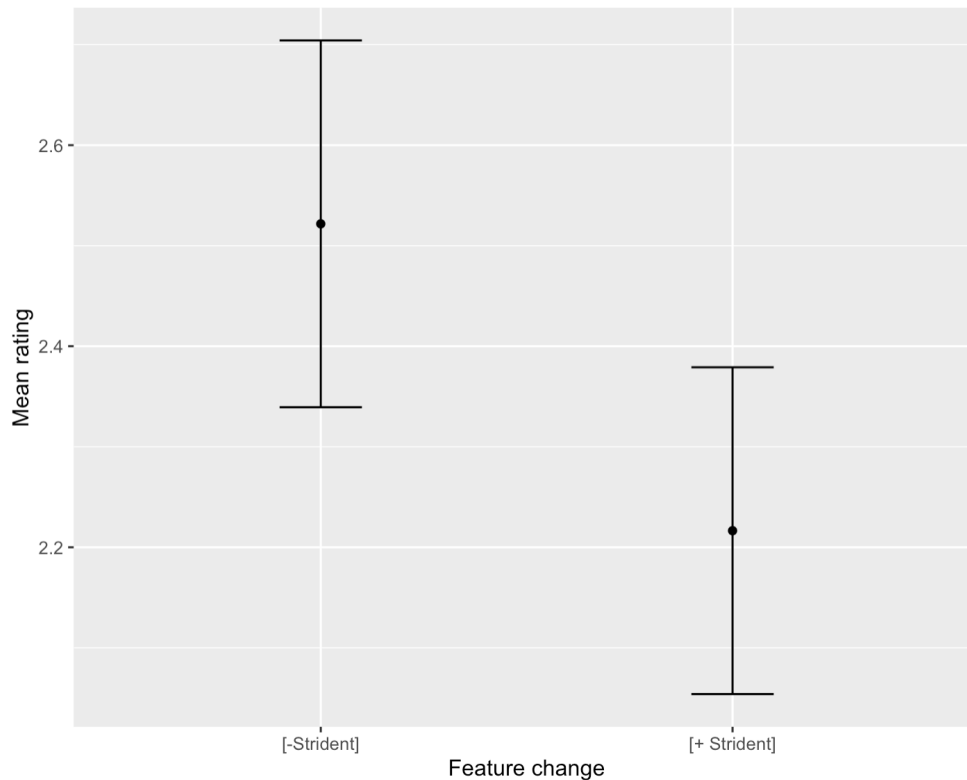


Figure 3.8: Mean ratings across fricatives. Error bars show a 95% confidence interval.

### 3.3 Discussion

Subjects did quite well at perceiving feature-based errors across the board and, even though they were not equally good at hearing the differences between each of the mispronunciation conditions and the control, their performances were fairly equivalent. Furthermore, the fact that all mispronunciation conditions showed both an increased likelihood of an incorrect response and a lower rating compared to the correctly pronounced control is evidence that voice and place errors were not so subtle that they were not being heard, and that manner errors, while different from the others, were not so different that they were at control. The central tendency measures for accuracy followed the same pattern as those for ratings across the different

conditions, with participants assigning the highest rating to the binomial expressions with no errors, the lowest rating to binomials with manner errors, and equivalent, slightly less bad, ratings to both place and manner errors.

This pattern was not observed, however, in my post-hoc analyses of the four inter-manner features and [+/-strident] fricatives. Within the manner feature, an increased likelihood of correctly recalling the target binomial corresponded with a higher mean rating of the phrase overall. This was somewhat unexpected, given the fact that the broader data show that a sound that is more disruptive to perception, which is to say perceived as a more egregious divergence from the expected sound, stands out more to a listener and is thus more accurately recalled. One possible explanation for these contrasting results might be that some mispronunciations are so unexpected and so disruptive that listeners find themselves basically unable to parse what they heard. This would explain how mispronunciations involving changes to the strident feature, which is characterized as being very noisy and therefore quite noticeable, are being recalled less accurately than changes in less noisy and perceptually weaker sounds. If I had more time I would like to rerun this experiment with a bigger set of materials that were distributed more evenly across the four manner features to look deeper into these results and to test whether the direction of the change had an effect on listener perception and accuracy.



## **4. Norming Study: Completing the couplet**

### **4.1 Background**

While the University of South Florida's collection of free association norms provided some background on the strength of the relationships between the conjuncts across many of my experimental binomial expressions, these data were collected during the years of 1973-1998, which is now between 25 to 50 years ago. That is quite a long time to assume that a pseudo-idiomatic expression will maintain its strength of association and frequency of usage among younger speakers, such as the college students who participated in my study (median age = 21). Because I was interested in whether expectation makes a listener more or less likely to process a sound error, it became necessary to probe how strong the relationship between the words that comprised each binomial was to a modern audience. To this end I ran an off-line norming study in which participants were asked to complete each of my original experiment's binomial expressions via free response.

### **4.2 Methods**

#### **4.2.1 Materials**

Items for the norming study consisted of one unfinished binomial for each of the experimental controls from Experiment 1, for a total of 60 unfinished binomials of the form X-and-\_\_\_\_, X-or-\_\_\_\_, and X-versus-\_\_\_\_.

### 4.2.2 Participants

43 individuals participated in the experiment with ages ranging from 18 to 42 ( $M = 29.5$ ,  $SD = 6.46$ ). Subjects were recruited from the experimenter's friend group and from an online Discord server. Four subjects were excluded for not responding to every question, and three subjects were excluded because they did not list English as one of the languages they spoke natively or learned before the age of six years old. I chose to exclude the latter group because it was impossible to confirm what level of familiarity these subjects had with English or the formulaic phrases being reviewed, given the fact that these subjects were recruited from a Discord server that the author is not a part of. This left data from 36 subjects for the analysis.

### 4.2.3 Procedure

The experiment was conducted online using Google Forms. Participants were asked to complete each of the unfinished phrases in the questionnaire with the first word that came to mind, based on the beginning of the phrase. They did this by typing their response into an empty textbox, as shown in Figure 4.1.

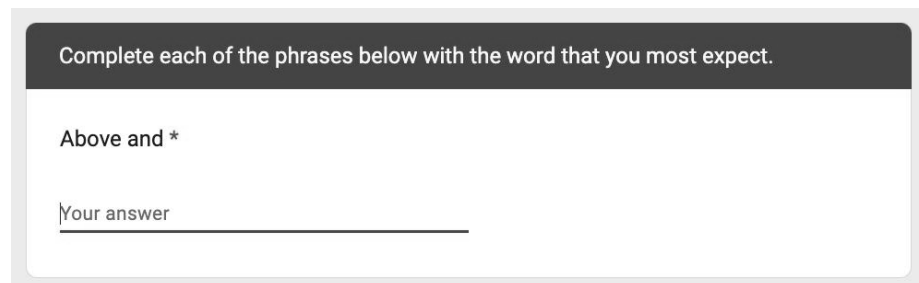
The image shows a sample item from a norming questionnaire. It consists of a dark grey header bar with the instruction "Complete each of the phrases below with the word that you most expect." Below the header, the text "Above and \*" is displayed. Underneath this text is a white text input field with a grey border and a horizontal line, containing the placeholder text "Your answer".

Figure 4.1: Sample item from norming questionnaire [Above and beyond]

### 4.3 Results

Responses from each of the 36 remaining subjects were standardized for spelling, number, and capitalization. Actual responses were compared against the expected binomial expressions used as stimuli in the main experiment. Participants' responses corresponded to the target binomial expressions from the main experiment between 3 and 36 times. Each of the binomial expressions was then labeled as one of three levels based on how often they were chosen by participants in the questionnaire: strongly associated, moderately associated, or weakly associated.

Binomial expressions were categorized as being weakly associated if participants more frequently chose a different word to complete the expression ( $n = 4$ ). Binomial expressions were categorized as strongly associated if participants chose the expected word more than 70% of the time ( $n = 45$ ). Finally, binomials that did not meet either of these criteria were categorized as moderately associated ( $n = 11$ ). Bar charts providing the responses for each of the items in the questionnaire can be found in Appendix C.

It is unsurprising, and even hoped for, given the nature of the original task, that the vast majority of the binomials in this experiment (75%) were regarded as strongly associated. I specifically chose to use binomial expressions as a type of formulaic speech (Carrol & Conklin, 2021) because of the strong associations between their component parts which allowed me to create expressions that would be predictable enough to allow listeners to generate expectations about what they were going to hear. Because of this, however, the distribution of binomial expressions

across the three association levels was very unbalanced, which made it difficult for any of the models I fit to the ratings and accuracy data to converge.

### 4.3.1 Accuracy by strength of association

The overall accuracy across strength of association and mispronunciation conditions is reported in Table 4.1. Participants were very good at correctly hearing and transcribing binomial expressions regardless of how strongly or weakly associated the final word in the binomial was with the rest of the phrase.

Table 4.1: Accuracy at different strengths of association

	Weak association		Moderate association		Strong association	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Control	22	0	58	1	219	2
Voicing	11	4	46	10	193	51
Place	17	5	43	11	186	51
Manner	22	3	49	12	198	31
Total (%)	72 (85.7)	12 (14.3)	196 (85.2)	34 (14.8)	796 (85.5)	135 (14.5)

A generalized linear mixed effects model was fit to the accuracy data with CONDITION, STRENGTH OF ASSOCIATION (SOA), and their interaction as fixed effects and random effects for SUBJECT and ITEM. There was a main effect of CONDITION, in line with the results in Section 3.2.1, but there was no effect of SOA, nor was there a significant interaction.

### 4.3.2 Ratings by strength of association

Figure 4.2 provides a graphical representation of the proportion of different ratings across conditions for strength of association and error type.

Impressionistically, it appears that an increase in a binomial expression’s SOA, which should reflect more highly activated anticipations about the final word, led to an increase in ratings of 3 and 4 across the mispronunciation conditions, and an increase in rating of 1 and 2 for control items.

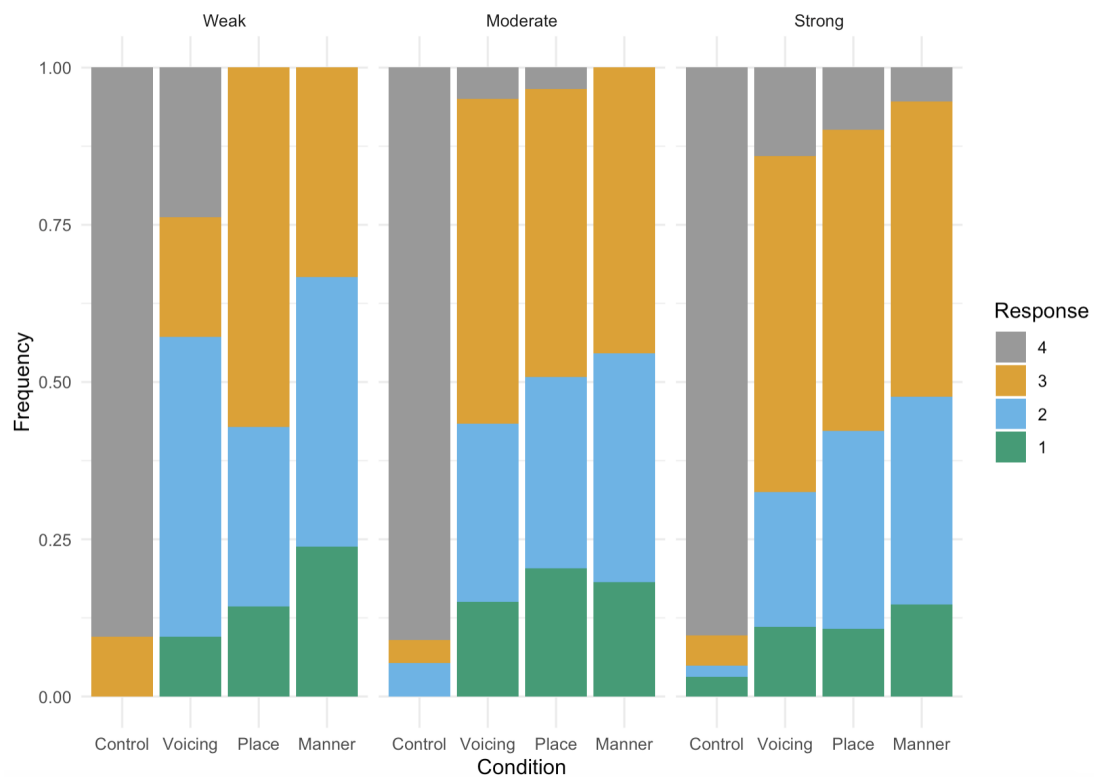


Figure 4.2: Effect of strength of association and type of feature change on pronunciation judgments

A cumulative link mixed effects model was fit to the ratings data with CONDITION, SOA, and their interaction as fixed effects, and SUBJECT and ITEM as

random effects. There was again a main effect of CONDITION, such that participants were more likely to rate a binomial expression with no mispronunciation higher than a binomial with a mispronunciation, but there was no effect of SOA, nor was there a significant interaction between CONDITION and SOA.

#### **4.4 Discussion**

The results of the models for accuracy and ratings preclude a simple main effect for strength of association; if it has an effect, it is subtle. As previously mentioned, the original binomial expressions that were chosen for this project were curated so as to include only binomials with fairly high levels of association between their conjuncts, which naturally led to a dramatically unbalanced dataset when trying to compare binomials with different association levels against each other. It is possible that this design choice, although necessary for the main experiment, led to this task not having enough data to reach significance for the current analysis.

In spite of not finding a significant effect of binomial strength, I did find some interesting trends in the ratings data. The more weakly associated the conjuncts in a binomial expression, the more they resemble everyday non-formulaic speech, and so I treated the weak association condition as the baseline. Based on the results of the norming study, listeners gave only correctly pronounced binomials and, to a lesser degree, items with mispronunciations in the voice feature the highest rating of 4 (*Completely as expected*) when those binomials were weakly associated. The fact that binomials with voicing errors in the target segments were the only mispronunciations

that were rated as high as the control condition supports my extension of Steriade's claim about voicing changes in word-final positions to voicing changes in word-initial positions. These ratings data show subjects being least perturbed by changes to the voicing feature, which supports my prediction, derived from Steriade's P-Map, that changing the voice feature of a sound is the most minimal feature change that one can make to a sound, regardless of context. These data also align with Martin & Pepperkamp's (2015) results, in which voicing mispronunciations were argued to be less detrimental to processing than manner and place mispronunciations. The fact that voicing errors are not flagged as poor pronunciations while place errors are provides further support for Lovitt & Allen's (2006) updated confusion matrices over Miller & Nicely's (1955).

The data showed more ratings of 3 (*Mostly as expected*) compared to 2 (*Mostly unexpected*) being assigned to items with voice and manner errors when a binomial's conjuncts were moderately associated. Ratings of 4 decreased in the voice condition, and 2 ratings increased for the control. And finally, when the conjuncts in a binomial had a high level of association, there was both a higher proportion of 4s and a lower proportion of 1s assigned to items in the manner and place conditions as well as a higher proportion of 3s and a lower proportion of 2s for voice and manner. Taken together, it does seem as though binomial expressions with more strongly associated conjuncts do cause listeners to be more forgiving of sound errors across all feature mispronunciations, but with an especially pronounced effect on manner errors. The trends that appear in this limited dataset provide support for the argument that there is

a continuum of predictability along which, at a certain point, the top-down expectation primed by a predictable environment starts to have more of an impact on a listener's processing of the signal, allowing more mispronunciations through.

Although I did not find a quantifiably supported effect of strength of association on listener accuracy or ratings judgments in this experiment, further research, with more evenly distributed lists for weak, moderate, and strong associations, must be done to truly rule it out.

## **5 General Discussion**

The ratings and transcriptions tasks in my main experiment were designed to examine the relative importance of different phonological features on the likelihood of recognizing sound errors. Both the ratings and transcriptions tasks found a main effect of error vs. no error, with subjects demonstrating lower accuracy in the recollection of errorful vs. error-free binomial expressions, and rating all mispronunciation conditions worse than correctly pronounced controls. Although accuracy was lower in the mispronunciation conditions, participants were quite good overall at accurately transcribing the phrases they heard across both control and experimental conditions. Pairwise comparisons were run for each of the three mispronunciation conditions (voice error, place of articulation error, and manner of articulation error), the results of which found a greater negative effect of manner errors on ratings compared to voice errors, and increased levels of accuracy when recalling items with manner feature errors versus items with voice and place feature



errors. No significant difference was found between voice and place errors for either experiment.

The results of my study were compared to the results of Martin & Pepperkamp's (2015) paper, *Asymmetries in the exploitation of phonetic features for word recognition*, which used a similar item design to test which articulatory feature classes listeners relied on more for word recognition. In Martin & Pepperkamp's study, subjects heard a list of items and were tasked with pressing a key every time they thought they heard a word and then typing the word they thought they had heard into a text box. Subjects were told that the list was being read by a stroke patient who had a high level of difficulty with producing intelligible words.

Martin & Pepperkamp found a significant difference between all three mispronunciation conditions and the control, as well as a significant difference between the voicing mispronunciation condition and the mispronunciation conditions for place and manner, with subjects correctly recognizing words with a modified voice feature more often than place and manner. Unlike my transcription task results, no difference between place and manner was observed. Also unlike my results, participant performance in Martin & Pepperkamp's mispronunciation conditions was very low, which Martin & Pepperkamp attribute to an increased difficulty in word recognition based on even a one-feature change. I wonder if this difference might be partially attributed to giving their participants the impression that the speaker would be an unreliable source for correctly pronounced words, leading them to assume the majority of the items were mispronounced, whereas my experiment gave no

qualifications about the speaker, perhaps leading my participants to assume the majority of the items would be correctly pronounced. The difference between these expectations may have set the bar for correct pronunciations higher and lower, respectively.

The overall shape of the results from my ratings task is in line with Martin & Pepperkamp's conclusion that voicing mispronunciations are less detrimental to understanding than place and manner mispronunciations, and are therefore less informative and important to word recognition. These results also provide support to the claim that changing the voice feature is a more minimal change than changing the place or manner of articulation features of a sound. Figure 3.4 (CIs for error ratings) in Section 3.2.2 shows a consistent increase in ratings values from manner to place to voice, such that subjects rated mispronunciations in the voice feature as being less unexpected- which can be thought of as being less detrimental- compared to manner pronunciations and, to a lesser extent, place mispronunciations. These results contradict Miller & Nicely's conclusion that voicing features are among the most robust, since that should mean changes in their features would be particularly salient, even more salient than changes to the different manner features. Participants tended to make more mistakes in my recall task on phrases with voicing mispronunciations compared to those with manner mispronunciations, which supports Martin & Pepperkamp's claim that voicing errors are less detrimental and less informative: if we don't often rely on differences in the voice feature for recognizing and distinguishing words, it's reasonable to think that we would accept a lot more

variation in that feature without actually hearing the error for what it was. This holds true for voice vs. manner but, contrary to Martin & Pepperkamp's results, my experiment found almost no difference in accuracy between voice and place, and the difference between their ratings also failed to reach significance.

I originally made the prediction, extended from facts about minimal feature changes built into Steriade's P-Map, that mispronouncing a voicing feature would be less disruptive to processing than a mispronunciation involving manner or place of articulation, even in word-initial contexts, and my results generally support this. Participants demonstrated lower accuracy in correctly identifying errors caused by a change to the voicing feature versus a change in manner, and gave them higher ratings when compared with both place and manner feature changes. Taken together, these data show that subjects are less likely to perceive voicing contrasts than manner contrasts even in the more prominent context of word onsets. In the context of the P-Map, these results tell us that changes in manner features are perceptually larger than voicing or place changes, based on the fact that they are more reliably distinguished from a correctly pronounced control than voicing in ratings tasks and than both voice and place changes in accuracy.

My finding that changes to the manner feature are more perceptible than voice changes in ratings and accuracy tasks is inconsistent with what would be expected based on the results of Miller & Nicely's confusion matrices (1955). In Miller & Nicely's experiment they found the voice and nasal features to be the most robust and least affected by noise and low-pass systems, generating a prediction that changes to

the voice feature in my experiment should be much more easily perceived than any other feature change except for nasality. On the opposite end of the spectrum, place features were found to be the weakest sound cues, more susceptible to confusions than any of Miller & Nicely's other feature changes. I predicted, based on these claims, that the results of the ratings and recall tasks in my experiment would show a strong bias towards more accurately hearing voice errors over place errors, and rating changes to the voice feature as significantly worse. Neither of these predictions were represented in my results. For example, the 95% confidence intervals for mean accuracy in the main experiment for place and voice errors almost completely overlap, and ratings data show place errors as actually having more of a negative impact on listener rating than voice errors. My results instead agreed with the predictions made by Lovitt & Allen's updated version of the Miller & Nicely experiment (2006), in which the place feature was more robust. This is fairly strong evidence against this particular takeaway from Miller & Nicely's confusion matrices.

## **6 Conclusion**

The main purpose of this study was to investigate how changes to the different phonological features of voicing, place of articulation, and manner of articulation affect listener perception of speech errors. I ran an experiment using binomial expressions as the primary stimuli to measure the effect that these different feature changes had on a participant's likelihood of correctly perceiving an error in a contextually predictable environment. Using ratings and recall tasks, I found evidence that changing a sound's manner feature has more of an effect on word recognition and

error perception than changing its place or voice feature. This adds to a growing body of research on how certain distinctive features are more important for word recognition (Martin & Pepperkamp, 2015), resistance to degraded listening environments (Miller & Nicely, 1955; Lovitt & Allen, 2006), and perceptibility of different features as they apply to correspondence constraint ranking in Optimality Theory (Steriade, 2001). This study also hoped to explore how prediction affects the likelihood of perceiving a sound error in a second experiment, but the experimental design demanded by the first experiment made this unachievable. Although the current study did not find a significant effect of strength of association on listener ratings or accuracy judgments, this was likely due to the extremely unbalanced stimuli categories. In the future I would like to rerun the strength of association models with more evenly distributed lists for weak, moderate, and strong associations in order to fully rule out the potential effect of predictability on perception.

## **6.1 Future Directions**

Due to the time and associated constraints of completing a Master's degree during a global pandemic (but I suspect this conundrum is true of every thesis and dissertation, regardless of era), there remain a number of experiments and data-analysis that I would like to conduct, building off of these results. As I mentioned in the section above, I would like to start by running an updated version of this experiment with counterbalanced lists for the different levels of association so as to more intentionally test the effect of predictability on listener performance with sound errors involving different feature changes.

One task I would like to spend more time on is in teasing apart the differences between manner features. To begin, I would like to rerun my experiment with a much bigger set of materials in which the four different manner features are equally represented to create a more robust and balanced dataset for analyzing some of the trends observed here. The lateral feature in particular was relatively absent from my dataset because voiced alveolar lateral approximants can only generate mispronunciations for errors in manner due to facts about the phoneme inventory of SAE. This was a problem for my original experiment, but would not be for an experiment aimed at comparing different manner feature errors.

I would also have liked to run some more focused analyses of my results data. One outstanding question is whether different mispronunciation conditions are more prone to specific types of transcription errors. For example: based on the fact that listeners treat voicing errors as less disruptive, are voicing errors autocorrected more often than manner errors, and are manner errors more often completely misunderstood? I have included a table providing some of this information in Appendix D, but again these are questions that can only really be answered by rerunning my experiment with more balanced experimental conditions.

I was surprised by the fact that participants performed better on accurately identifying items with mispronunciations involving non-strident fricatives, since stridents are generally considered some of the noisiest consonants due to the rapid, turbulent airflow that is a cornerstone of the articulation. This was especially surprising considering the results of the main experiment, which found that manner

was generally assigned the lowest ratings but was also the most accurately recalled. I suspect the difference between the results of the main experiment and the results of the strident analysis has to do with the type of transcription error being made. The majority of incorrect transcriptions in the main experiment involved autocorrecting to the sound listeners expected to hear based on the context generated by the binomial expression, but I would predict that mistranscriptions involving strident sounds fall into one of the other categories in which the error leads to a breakdown in understanding of the word or phrase as a whole (See Tables D.2 and D.3 in Appendix D for a preliminary analysis of error type within manner and fricative conditions).

One final area of lingering curiosity is whether there is an effect of the direction of change on listener performance when processing different manner feature errors: I don't have any specific predictions about whether it is more jarring for a listener to hear a strident, for example, when they are expecting something else or to hear a different sound when expecting a strident, but I suspect that they probably are different.

## Appendix A: Experimental Stimuli

Item	Condition	Item
1	Control	arts and sciences
1	Voice error	arts and ziences
1	Place error	arts and fiences
1	Manner error	arts and tiences
2	Control	bread and butter
2	Voice error	bread and putter
2	Place error	bread and tutter
2	Manner error	bread and mutter
3	Control	brother and sister
3	Voice error	brother and zister
3	Place error	brother and hister
3	Manner error	brother and tister
4	Control	buy and sell
4	Voice error	buy and zell
4	Place error	buy and θell
4	Manner error	buy and tell
5	Control	supply and demand
5	Voice error	supply and temand
5	Place error	supply and gemand
5	Manner error	supply and lemand
6	Control	research and development
6	Voice error	research and tevelopment
6	Place error	research and bevelopment
6	Manner error	research and revelopment
7	Control	heart and soul
7	Voice error	heart and zoul
7	Place error	heart and θoul
7	Manner error	heart and toul
8	Control	pain and suffering
8	Voice error	pain and zuffering
8	Place error	pain and fuffering
8	Manner error	pain and tuffering
9	Control	safe and sound
9	Voice error	safe and zound
9	Place error	safe and θound



9	Manner error	safe and tound
10	Control	sweet and sour
10	Voice error	sweet and zour
10	Place error	sweet and four
10	Manner error	sweet and tour
11	Control	above and beyond
11	Voice error	above and peyond
11	Place error	above and deyond
11	Manner error	above and meyond
12	Control	add and subtract
12	Voice error	add and zubtract
12	Place error	add and hubtract
12	Manner error	add and tubtract
13	Control	gin and tonic
13	Voice error	gin and donic
13	Place error	gin and ponic
13	Manner error	gin and sonic
14	Control	show and tell
14	Voice error	show and dell
14	Place error	show and kell
14	Manner error	show and sell
15	Control	cats and dogs
15	Voice error	cats and togs
15	Place error	cats and gogs
15	Manner error	cats and rogs
16	Control	checks and balances
16	Voice error	checks and palances
16	Place error	checks and dalances
16	Manner error	checks and malances
17	Control	cloak and dagger
17	Voice error	cloak and tagger
17	Place error	cloak and bagger
17	Manner error	cloak and zagger
18	Control	song and dance
18	Voice error	song and tance
18	Place error	song and gance
18	Manner error	song and rance
19	Control	truth or dare

19	Voice error	truth or tare
19	Place error	truth or gare
19	Manner error	truth or zare
20	Control	coffee or tea
20	Voice error	coffee or dea
20	Place error	coffee or kea
20	Manner error	coffee or sea
21	Control	up or down
21	Voice error	up or town
21	Place error	up or bown
21	Manner error	up or lown
22	Control	live or die
22	Voice error	live or tie
22	Place error	live or bie
22	Manner error	live or zie
23	Control	analog or digital
23	Voice error	analog or tigital
23	Place error	analog or gigital
23	Manner error	analog or rigital
24	Control	good or bad
24	Voice error	good or pad
24	Place error	good or gad
24	Manner error	good or vad
25	Control	backwards and forwards
25	Voice error	backwards and vorwards
25	Place error	backwards and sorwards
25	Manner error	backwards and porwards
26	Control	crime and punishment
26	Voice error	crime and bunishment
26	Place error	crime and tunishment
26	Manner error	crime and funishment
27	Control	friends and family
27	Voice error	friends and vamily
27	Place error	friends and hamily
27	Manner error	friends and pamily
28	Control	flora and fauna
28	Voice error	flora and vauna
28	Place error	flora and hauna

28	Manner error	flora and pauna
29	Control	intents and purposes
29	Voice error	intents and burposes
29	Place error	intents and kurposes
29	Manner error	intents and furposes
30	Control	war and peace
30	Voice error	war and beace
30	Place error	war and teace
30	Manner error	war and feace
31	Control	brick and mortar
31	Voice error	brick and portar
31	Place error	brick and nortar
31	Manner error	brick and bortar
32	Control	life and death
32	Voice error	life and teth
32	Place error	life and geth
32	Manner error	life and leth
33	Control	tar and feather
33	Voice error	tar and veather
33	Place error	tar and feather
33	Manner error	tar and peather
34	Control	true or false
34	Voice error	true or valse
34	Place error	true or halse
34	Manner error	true or palse
35	Control	rise and fine
35	Voice error	rise and zine
35	Place error	rise and θine
35	Manner error	rise and tine
36	Control	cream and fugar
36	Voice error	cream and zugar
36	Place error	cream and fugar
36	Manner error	cream and tugar
37	Control	silver and gold
37	Voice error	silver and kold
37	Place error	silver and dold
37	Manner error	silver and hold
38	Control	divide and conquer

38	Voice error	divide and gonquer
38	Place error	divide and ponquer
38	Manner error	divide and honquer
39	Control	horse and carriage
39	Voice error	horse and garriage
39	Place error	horse and tarriage
39	Manner error	horse and harriage
40	Control	hugs and kisses
40	Voice error	hugs and gisses
40	Place error	hugs and tisses
40	Manner error	hugs and hisses
41	Control	this or ðat
41	Voice error	this or θat
41	Place error	this or zat
41	Manner error	this or nat
42	Control	hot or cold
42	Voice error	hot or gold
42	Place error	hot or pold
42	Manner error	hot or hold
43	Control	us versus ðem
43	Voice error	us versus θem
43	Place error	us versus zem
43	Manner error	us versus nem
44	Control	name and number
44	Voice error	name and tumber
44	Place error	name and mumber
44	Manner error	name and rumber
45	Control	day and night
45	Voice error	day and tight
45	Place error	day and might
45	Manner error	day and dight
46	Control	fork and (k)nife
46	Voice error	fork and tife
46	Place error	fork and mife
46	Manner error	fork and zife
47	Control	newspapers and magazines
47	Voice error	newspapers and pagazines
47	Place error	newspapers and nagazines

47	Manner error	newspapers and bagazines
48	Control	mix and match
48	Voice error	mix and patch
48	Place error	mix and natch
48	Manner error	mix and batch
49	Control	nature versus nurture
49	Voice error	nature versus turture
49	Place error	nature versus murture
49	Manner error	nature versus zurture
50	Control	hammer and nail
50	Voice error	hammer and tail
50	Place error	hammer and mail
50	Manner error	hammer and lail
51	Control	television or radio
51	Voice error	television or tadio
51	Place error	television or jadio
51	Manner error	television or nadio
52	Control	action versus reaction
52	Voice error	action versus teaction
52	Place error	action versus jeaction
52	Manner error	action versus zeaction
53	Control	left or right
53	Voice error	left or tight
53	Place error	left or jight
53	Manner error	left or dight
54	Control	expectation versus reality
54	Voice error	expectation versus teality
54	Place error	expectation versus jeality
54	Manner error	expectation versus neality
55	Control	trick or treat
55	Voice error	trick or dreat
55	Place error	trick or preat
55	Manner error	trick or freat
56	Control	paper or plastic
56	Voice error	paper or blastic
56	Place error	paper or klastic
56	Manner error	paper or flastic
57	Control	fight or flight

57	Voice error	fight or vlight
57	Place error	fight or slight
57	Manner error	fight or plight
58	Control	eat or drink
58	Voice error	eat or trink
58	Place error	eat or grink
58	Manner error	eat or zrink
59	Control	past or present
59	Voice error	past or bresent
59	Place error	past or tresent
59	Manner error	past or fresent
60	Control	business or pleasure
60	Voice error	business or bleasure
60	Place error	business or kleasure
60	Manner error	business or fleasure

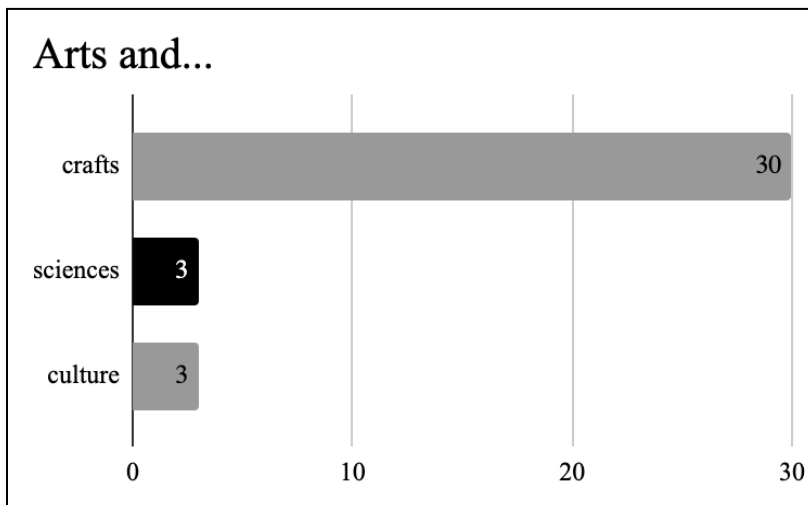
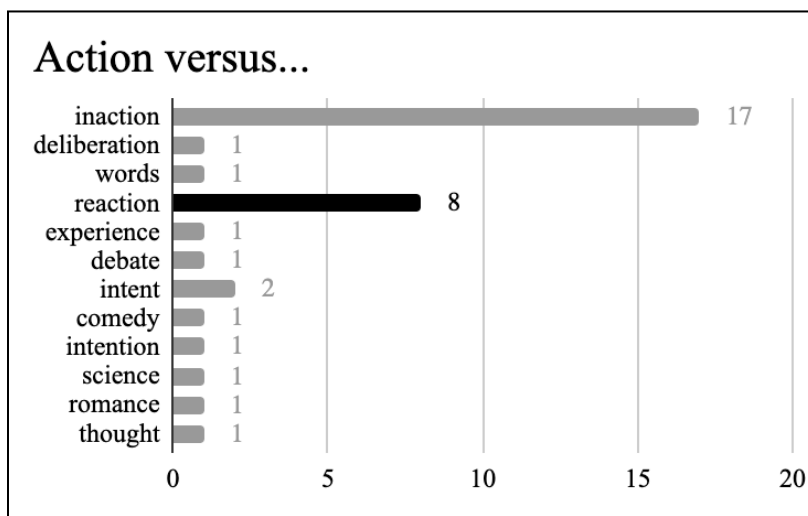
## Appendix B: Manipulations by phoneme

Base	Voicing	Place	Manner
[d] = 10	[t] x 10	[g] x 6; [b] x 4	[l] x 3; [ɹ] x 4; [z] x 3
[b] = 4	[p] x 4	[d] x 3; [g] x 1	[m] x 3; [v] x 1
[n] = 5	[t] x 5	[m] x 5	[d] x 1; [z] x 2; [l] x 1; [ɹ] x 1
[ɹ] = 4	[t] x 4	[j] x 4	[n] x 2; [z] x 1; [d] x 1
[s] = 8	[z] x 8	[f] x 3; [h] x 2; [θ] x 3	[t] x 8
[f] = 5	[v] x 5	[s] x 1; [h] x 3; [ʃ] x 1	[p] x 5
[ʃ] = 2	[ʒ] x 8	[f] x 1; [θ] x 1	[t] x 2
[t] = 3	[d] x 3	[p] x 1; [k] x 2	[s] x 3
[k] = 4	[g] x 4	[p] x 2; [t] x 2	[h] x 4
[p] = 3	[b] x 3	[t] x 2; [k] x 1	[f] x 3
[g] = 1	[k] x 1	[d] x 1	[h] x 1
[m] = 3	[p] x 3	[n] x 3	[b] x 3
[ð] = 2	[θ] x 2	[z] x 2	[n] x 2
[tɹ] = 1	[dɹ] x 1	[pɹ] x 1	[sɹ] x 1
[dɹ] = 1	[tɹ] x 1	[gɹ] x 1	[zɹ] x 1
[fl] = 1	[vl] x 1	[sl] x 1	[pl] x 1
[pl] = 2	[bl] x 2	[kl] x 2	[fl] x 2
[pɹ] = 1	[bɹ] x 1	[tɹ] x 1	[fɹ] x 1

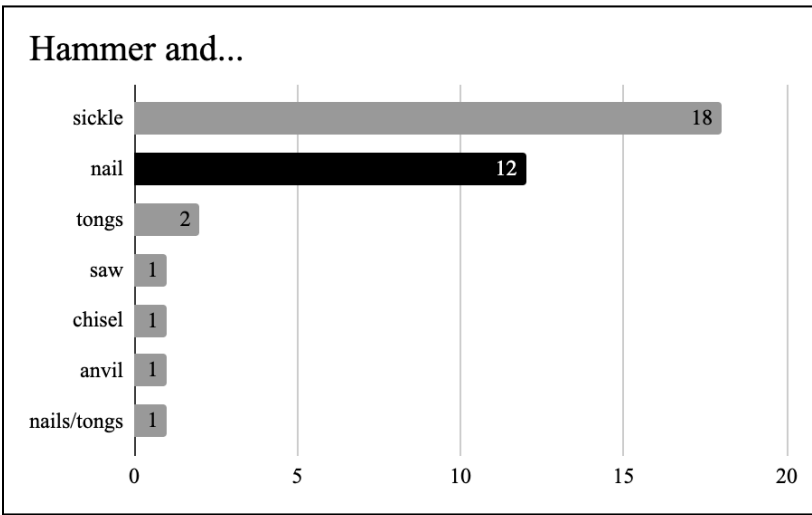
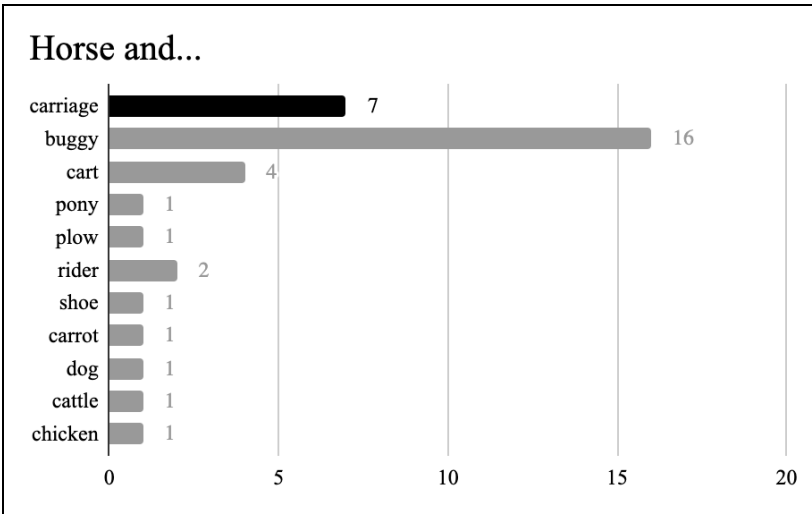
## Appendix C: Results from Norming Study

Note that words that correspond with the second half of the binomial expression in the main experiment are represented by the darker colored bars.

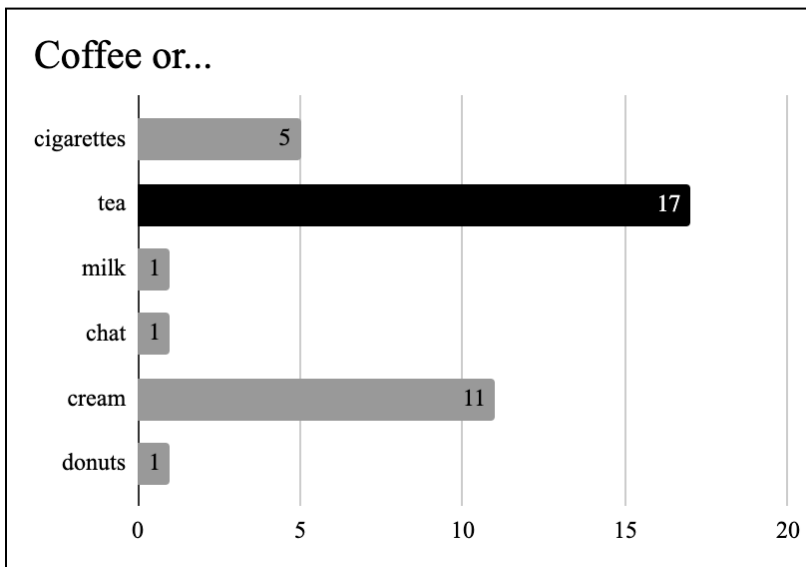
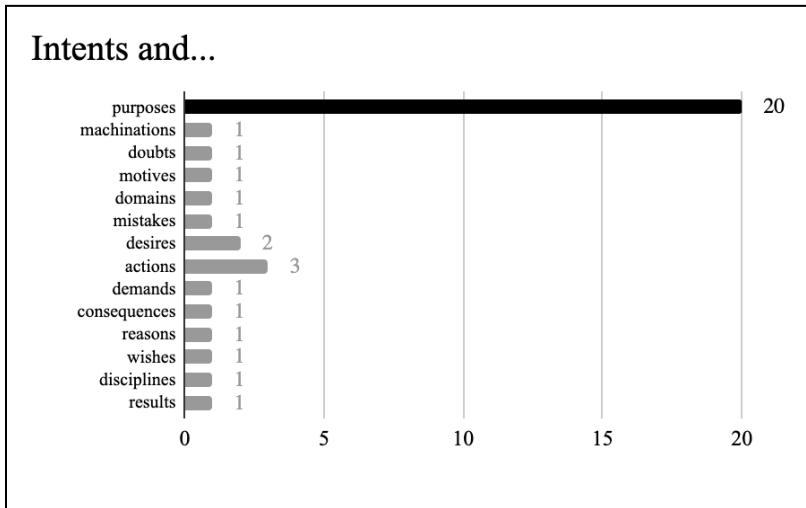
### C.1 Weakly associated binomials

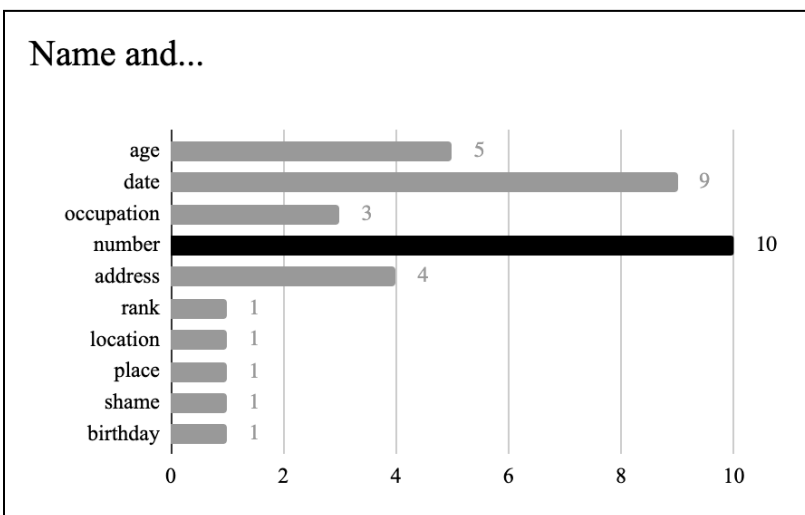
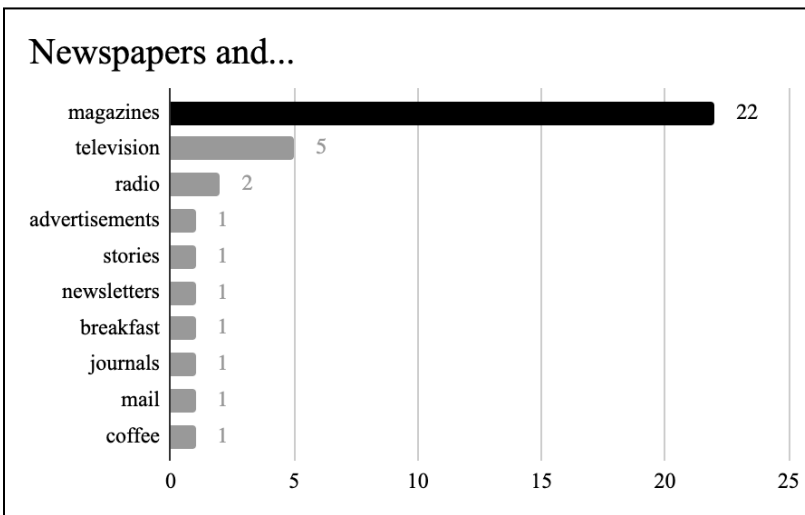
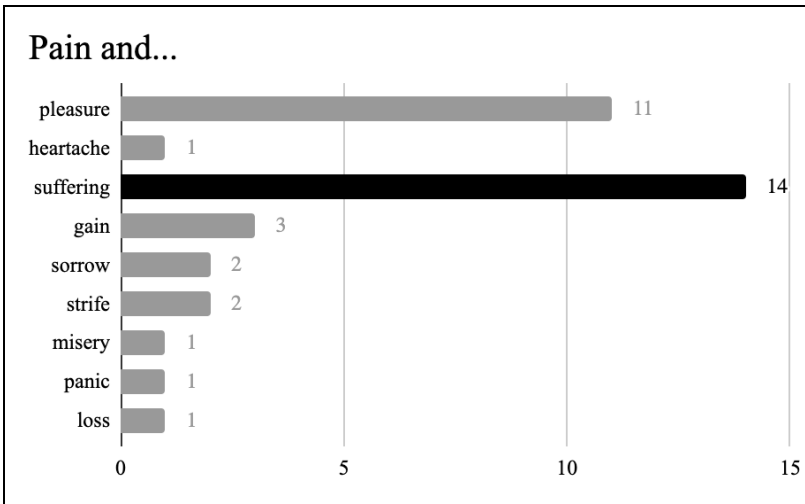


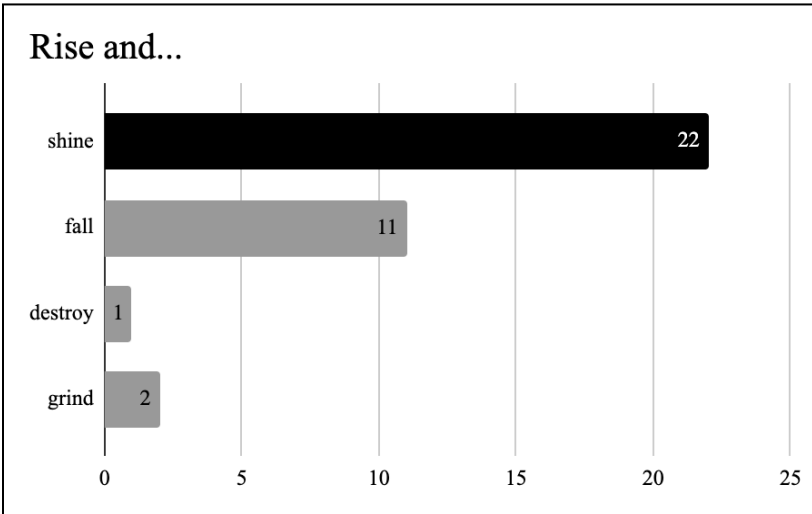
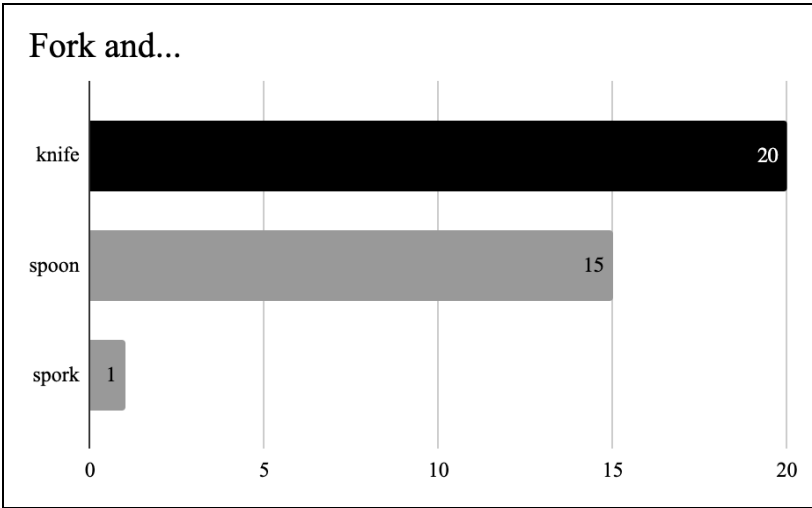


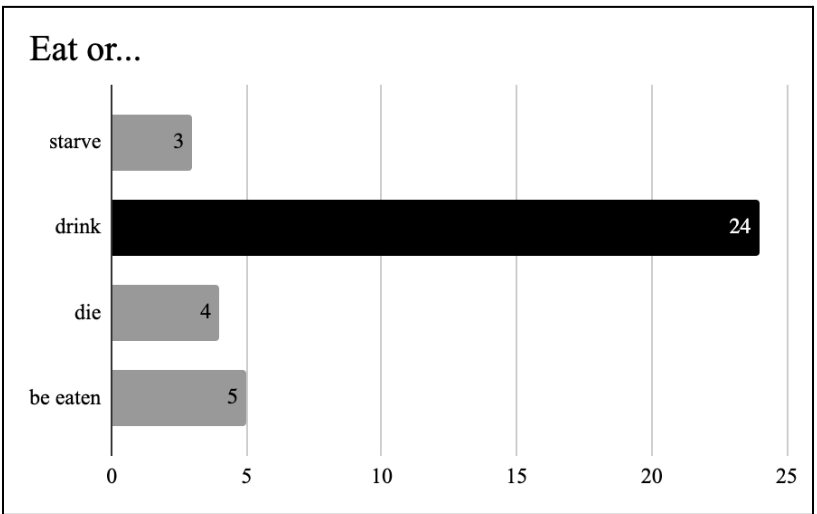
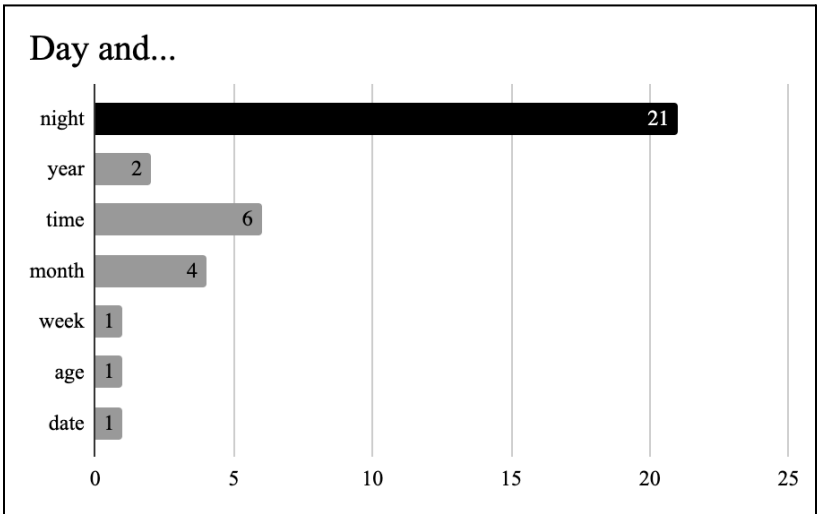
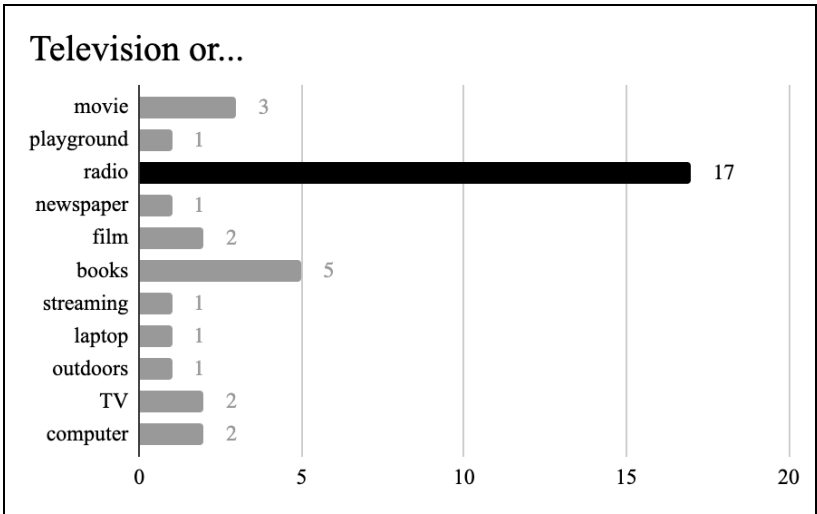


## C.2 Moderately associated binomials

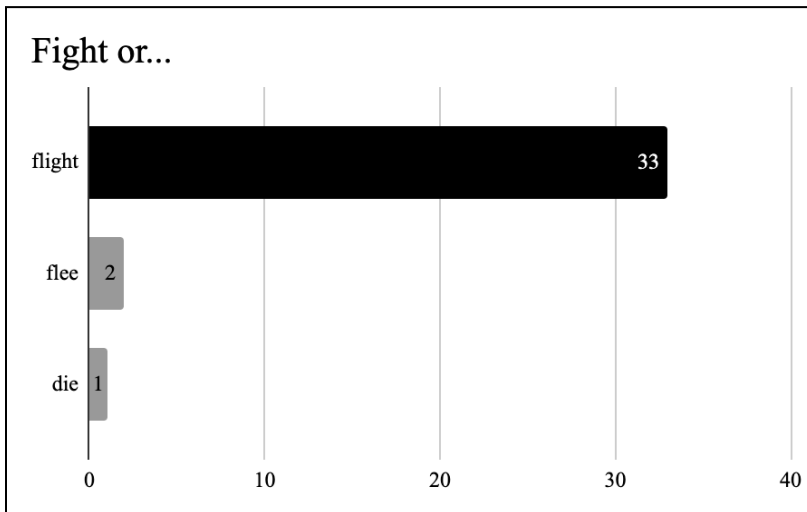
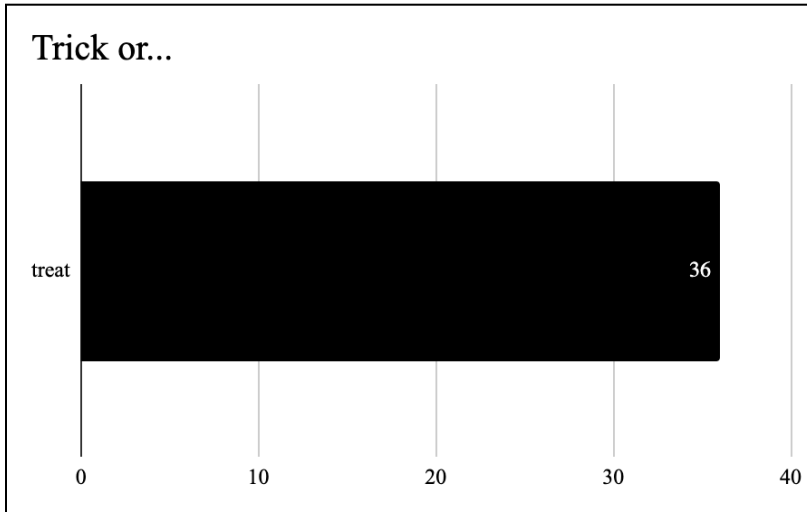


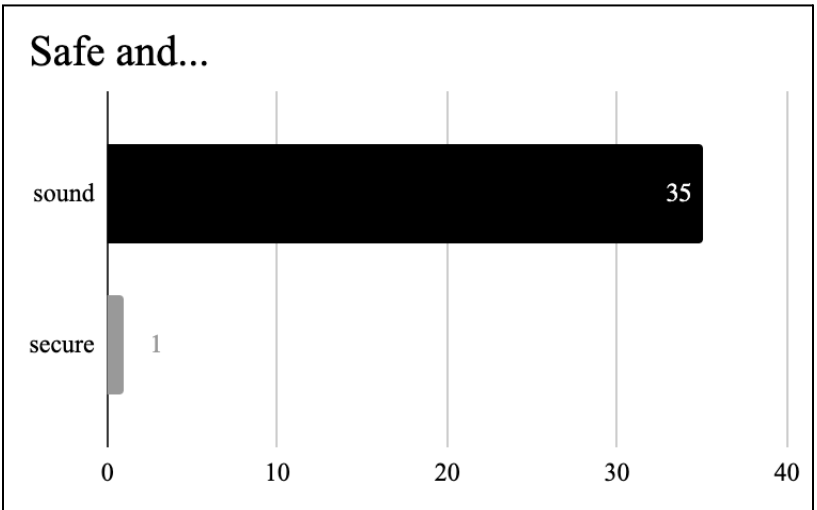
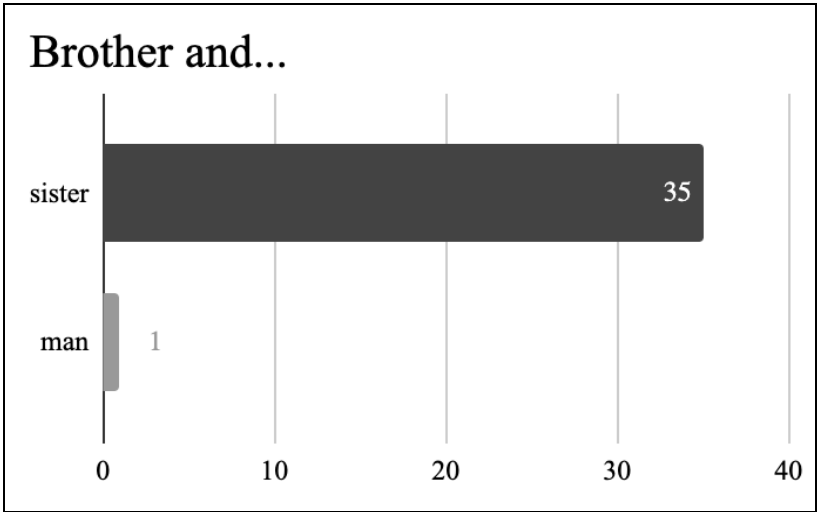


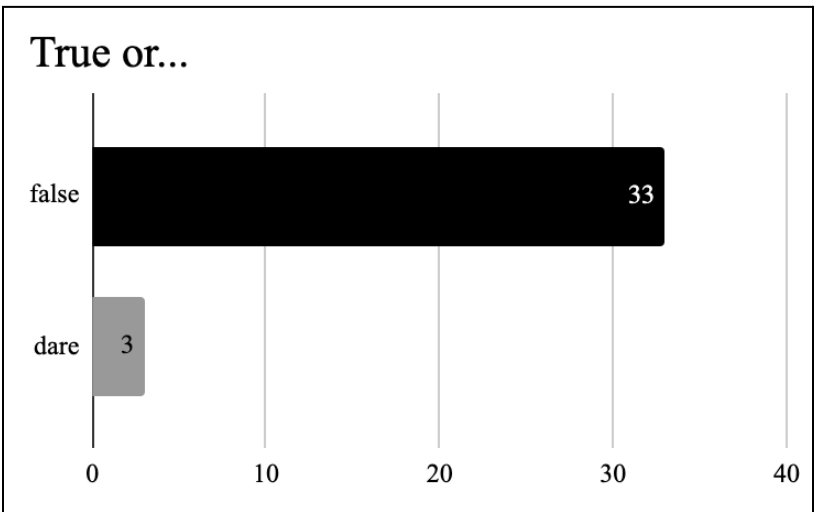
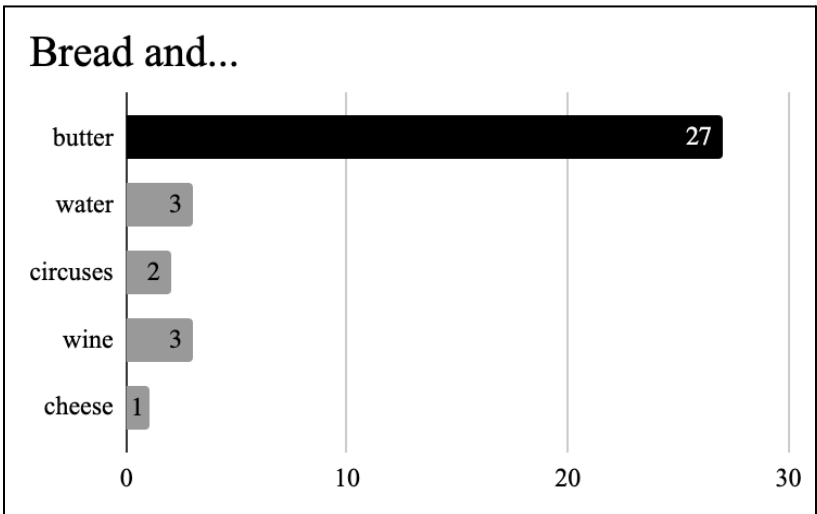
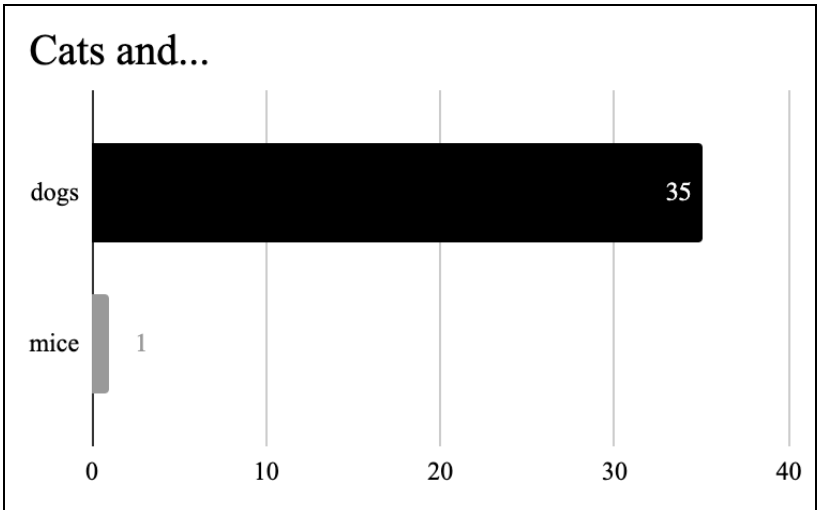




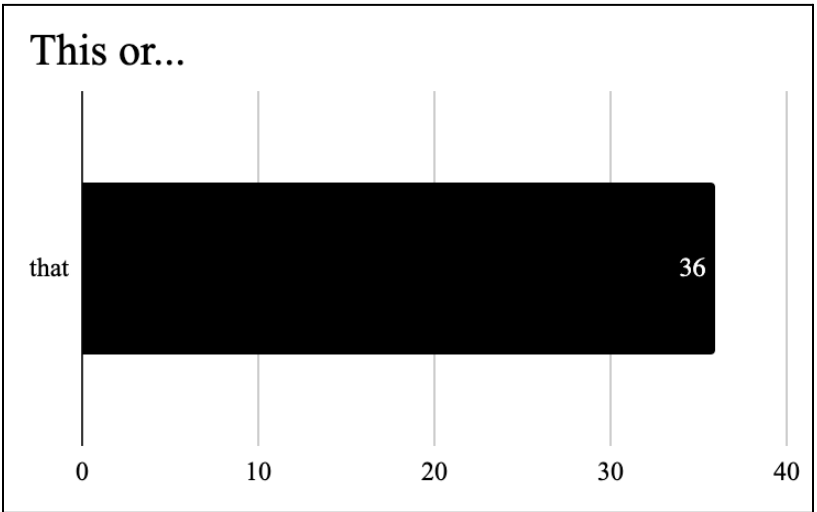
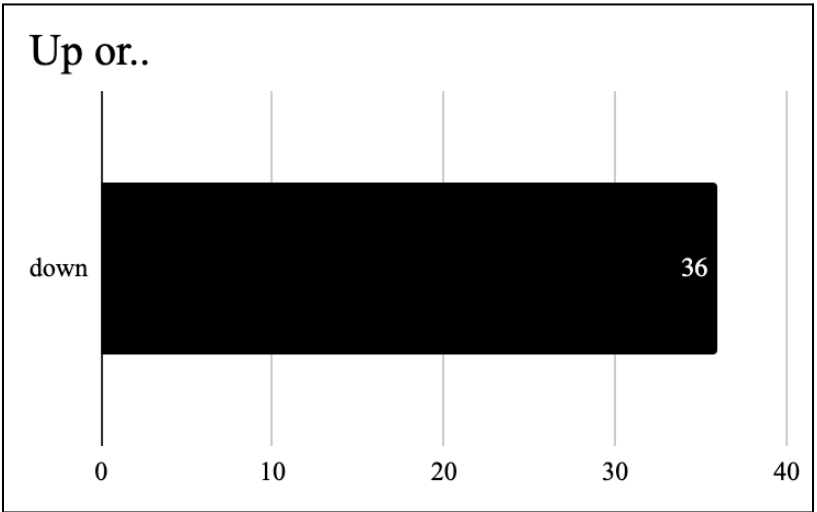
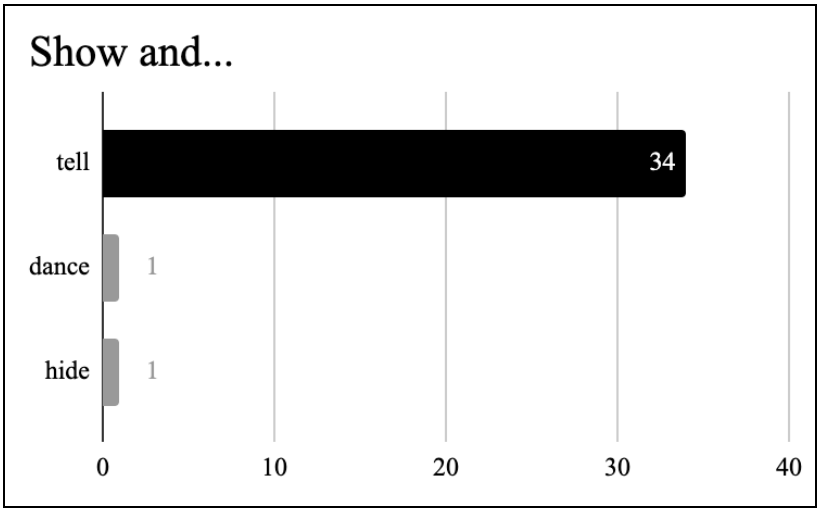
### C.3 Strongly associated binomials

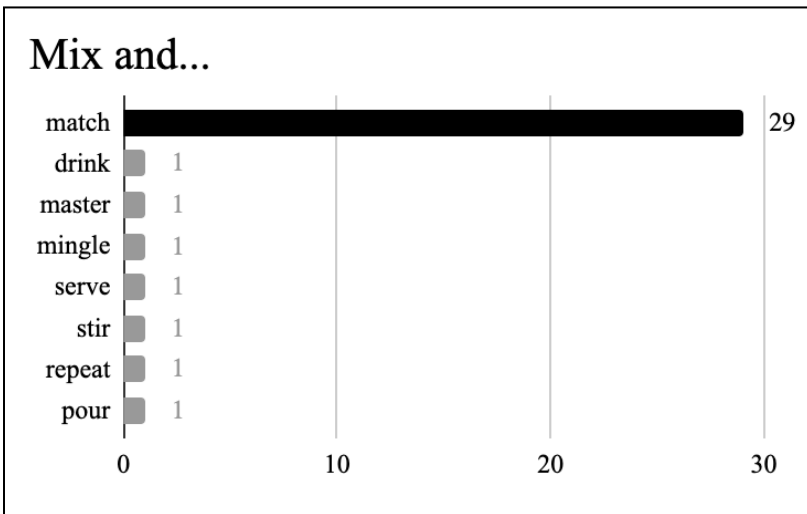
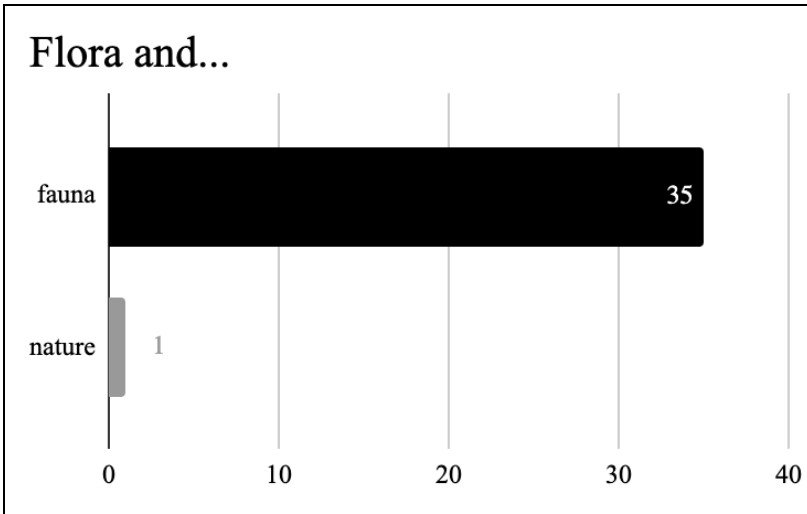


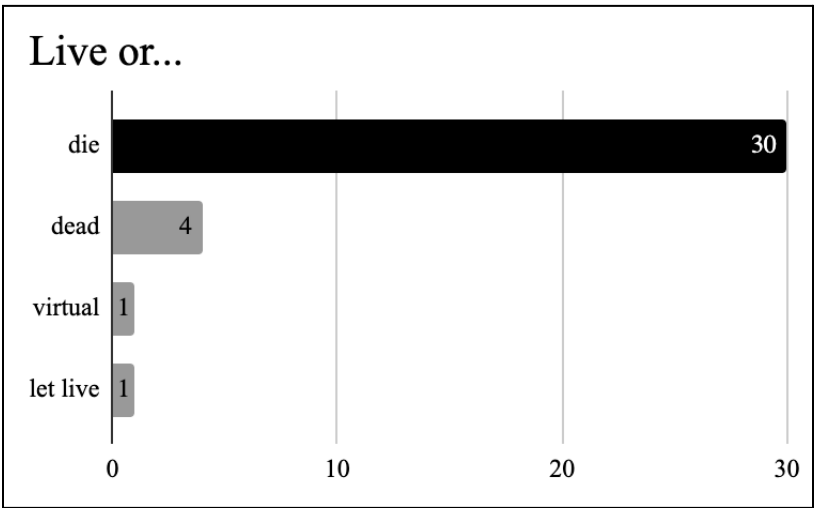
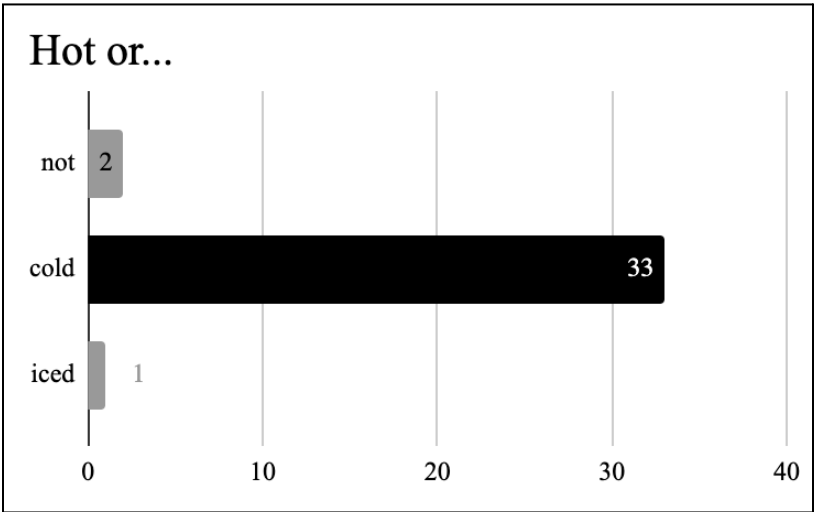
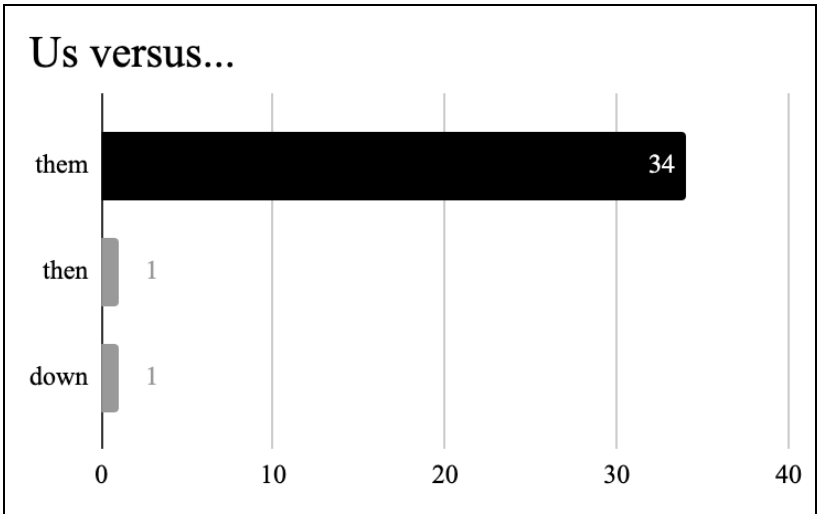


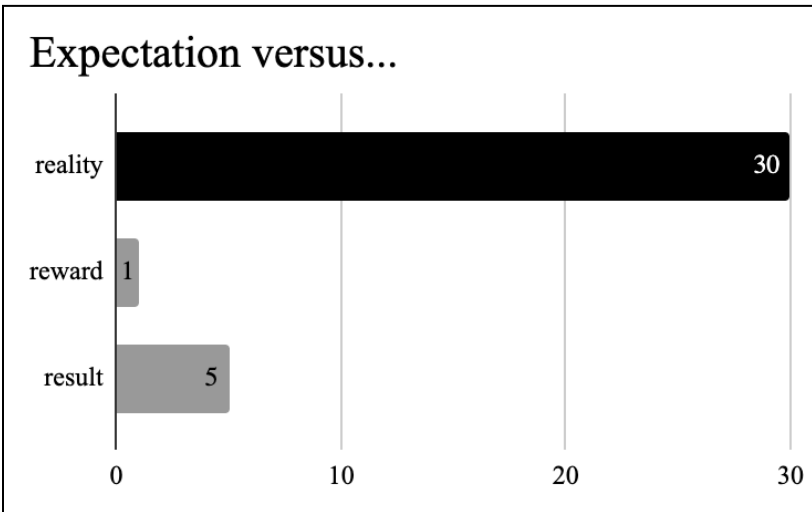
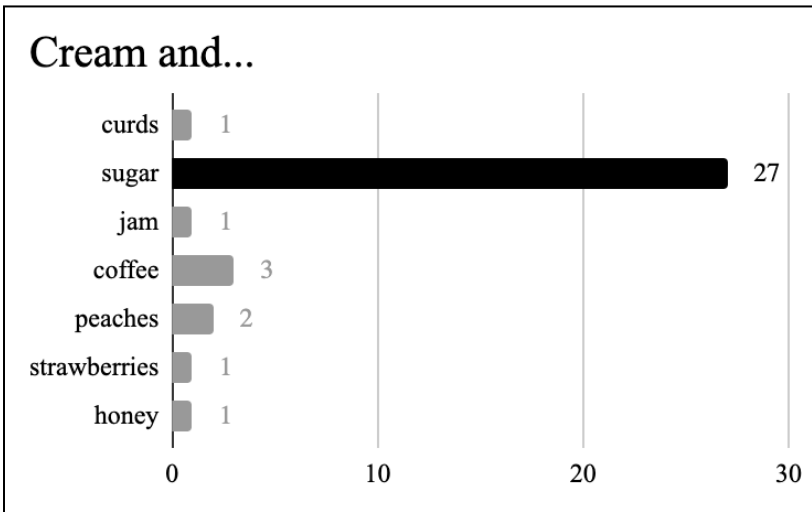
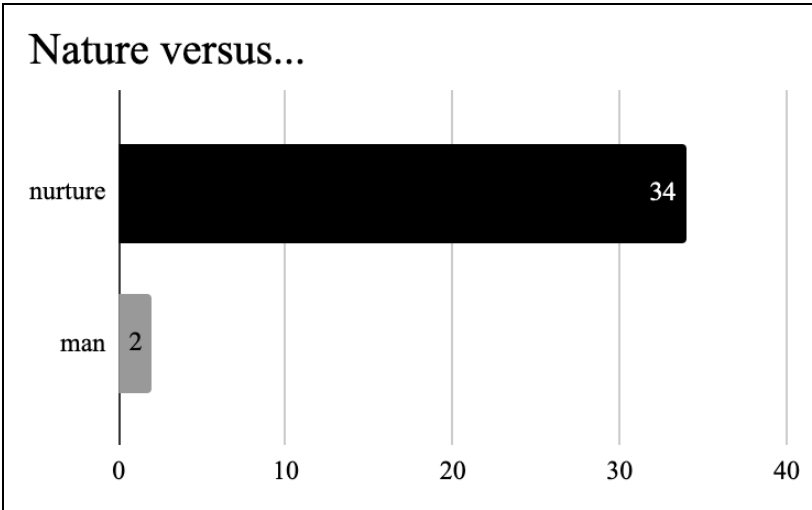


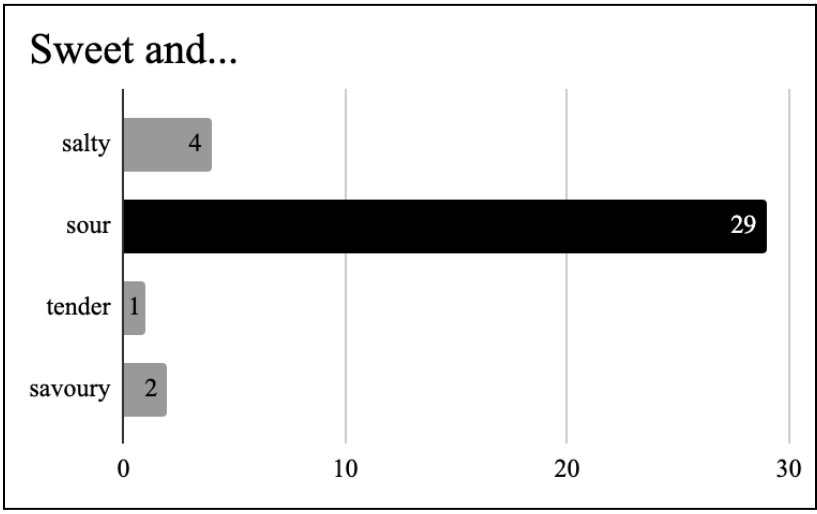
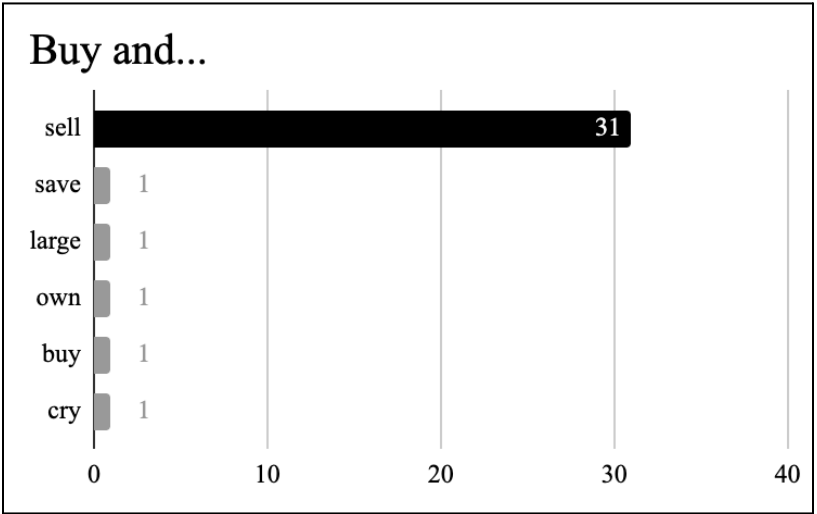
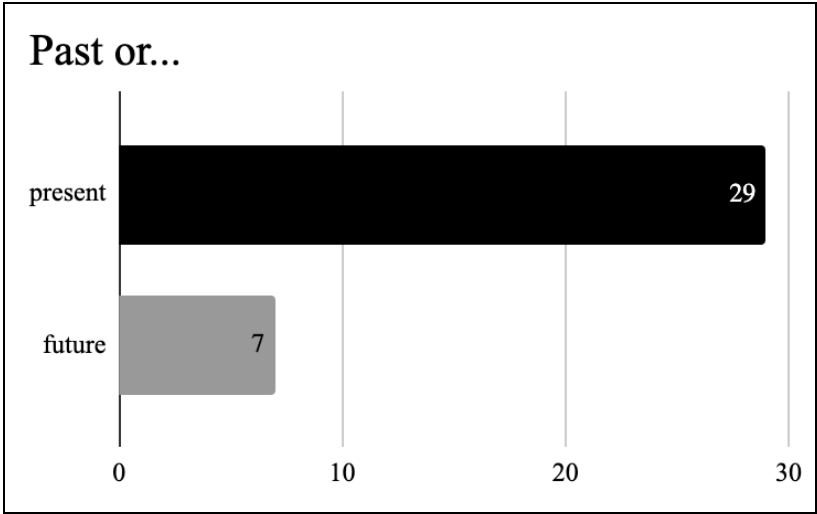


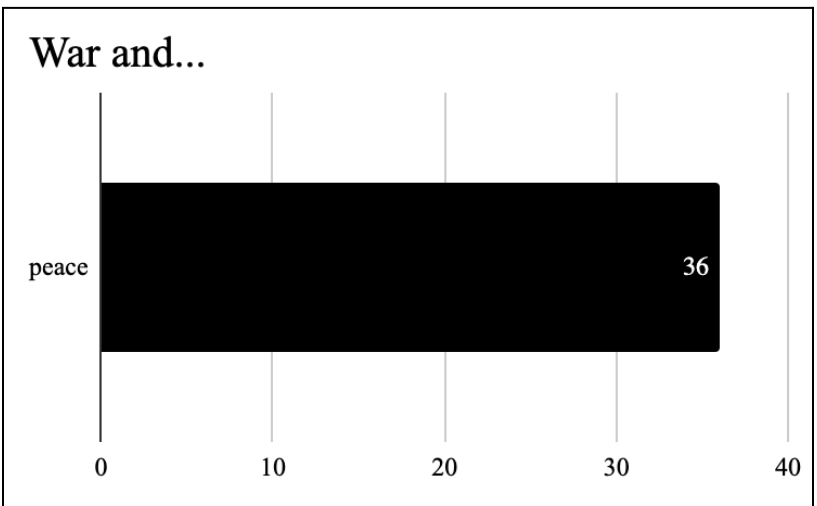
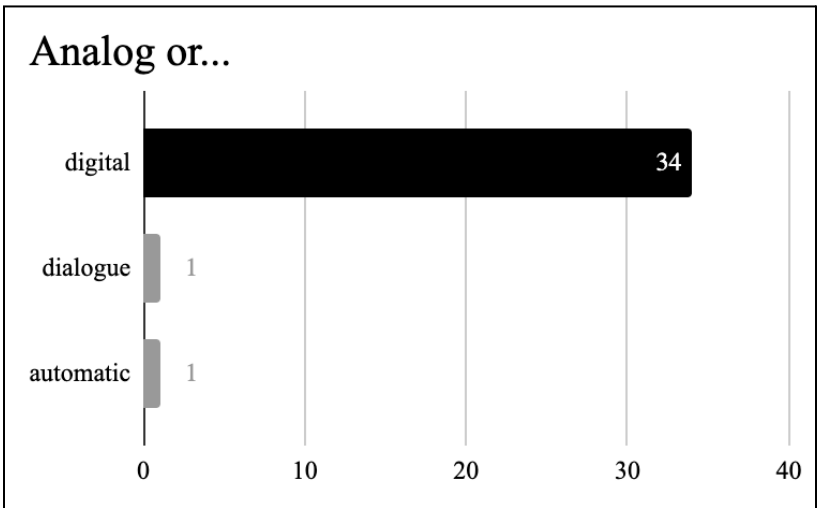
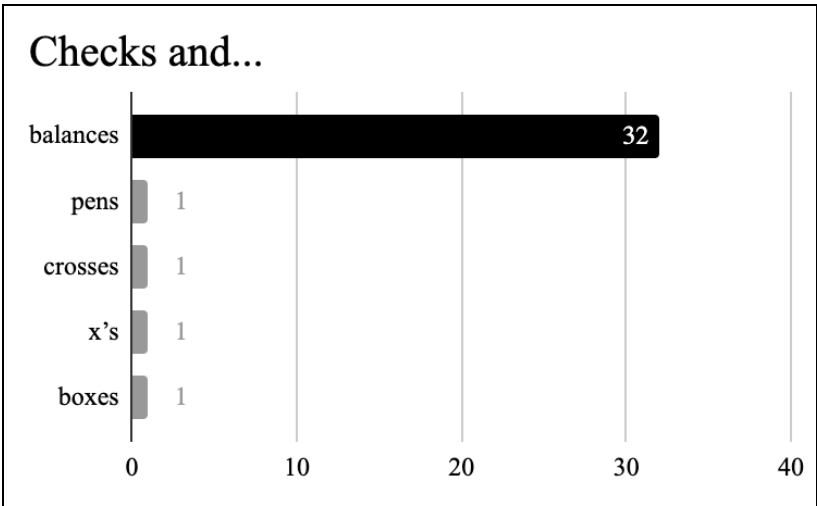


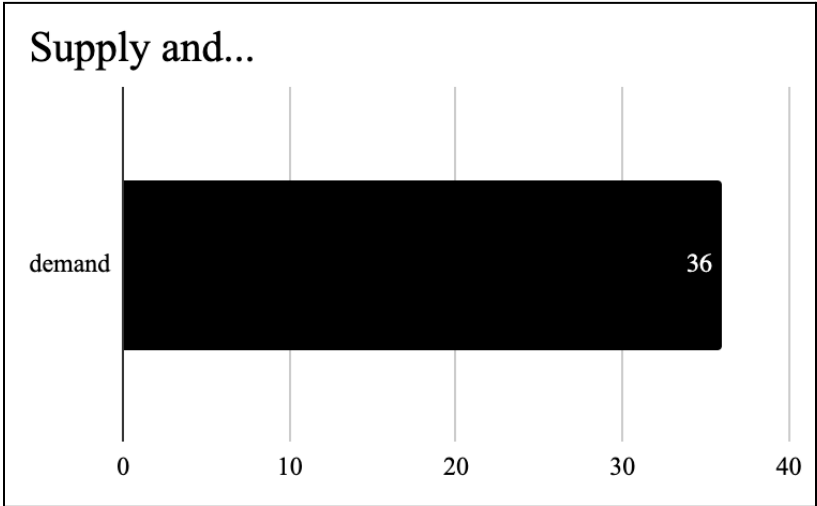
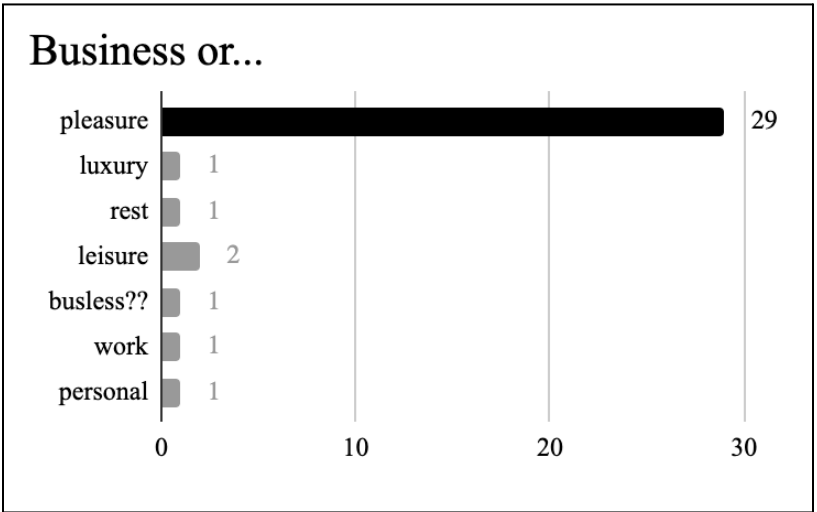
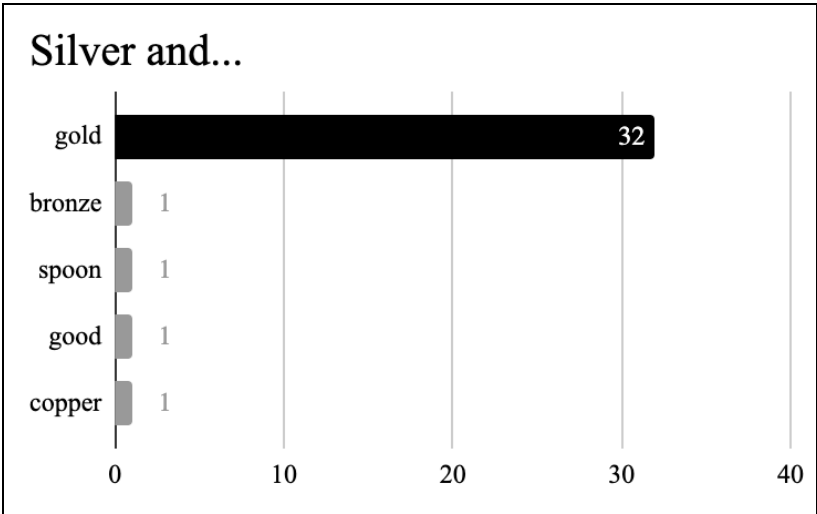


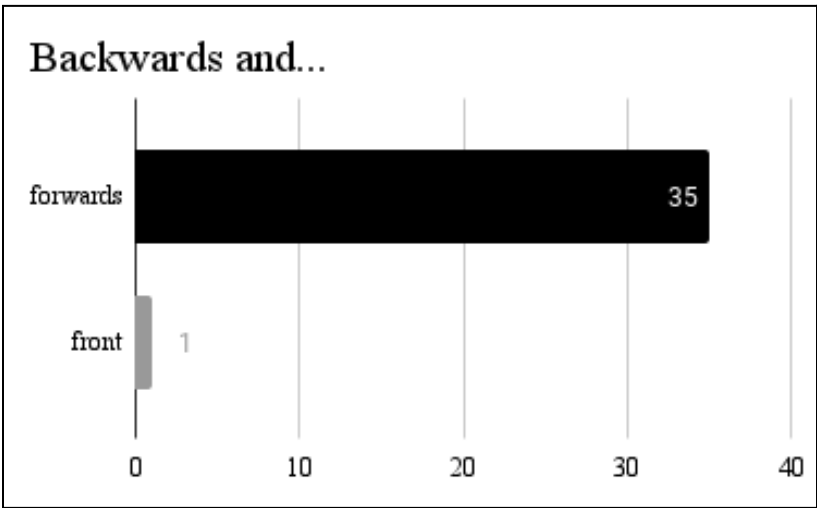
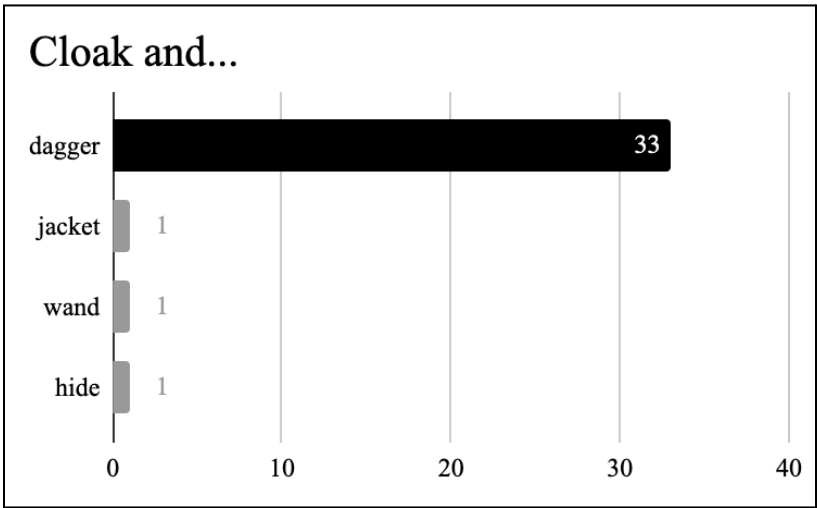
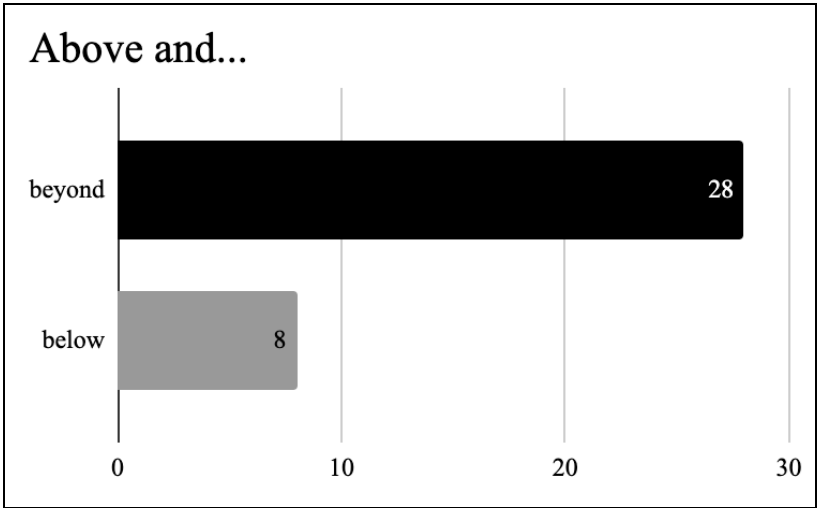




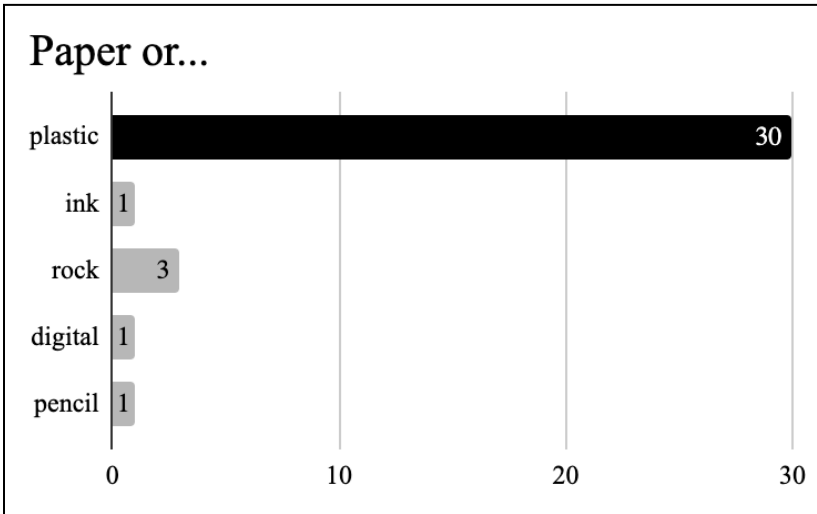
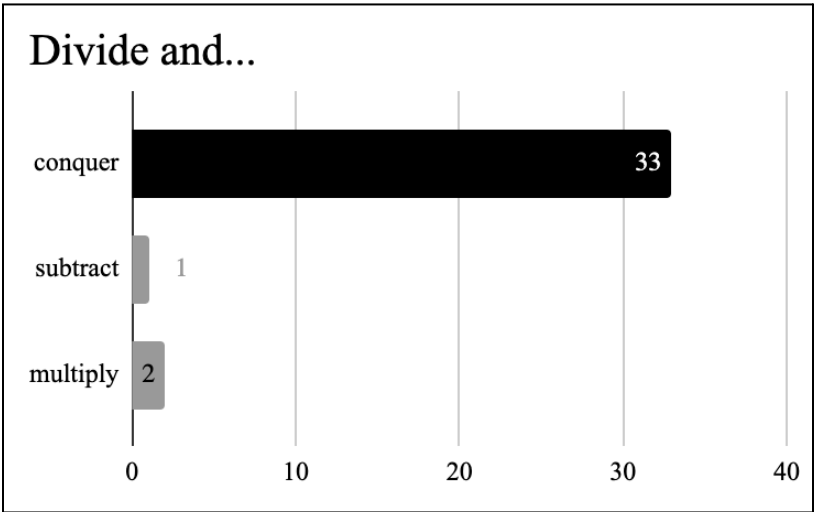
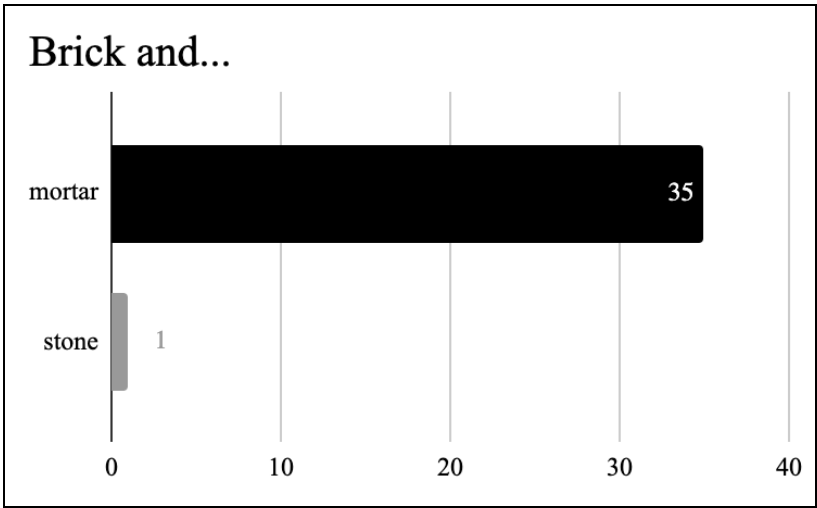


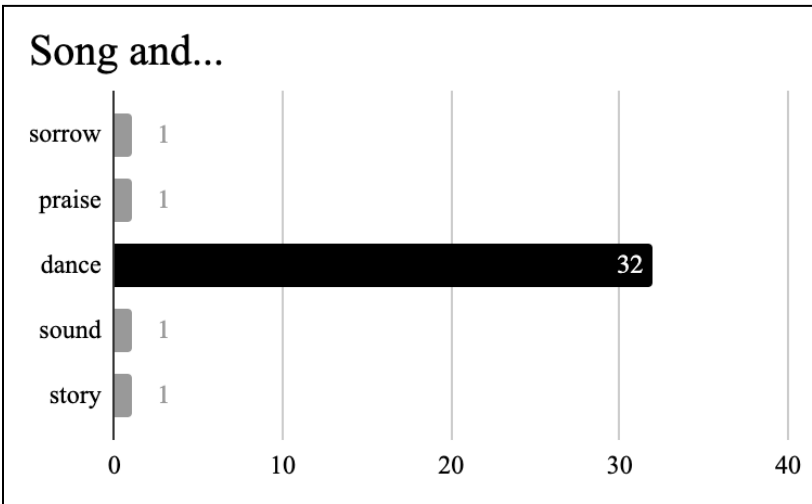
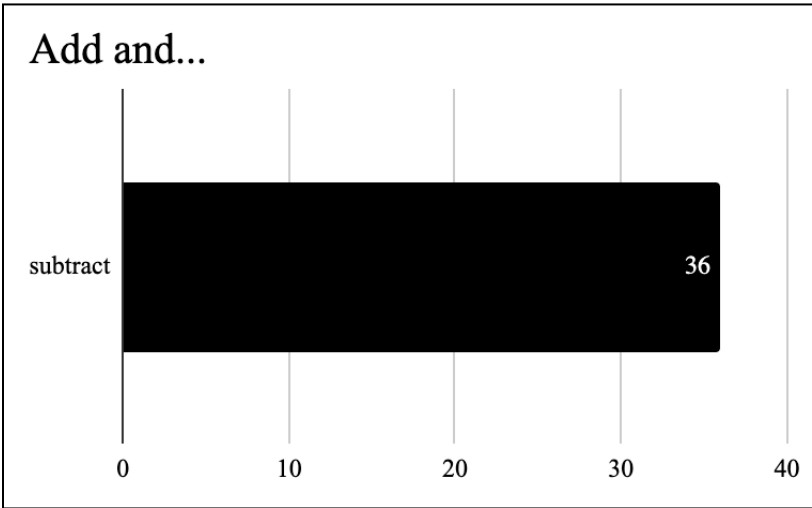
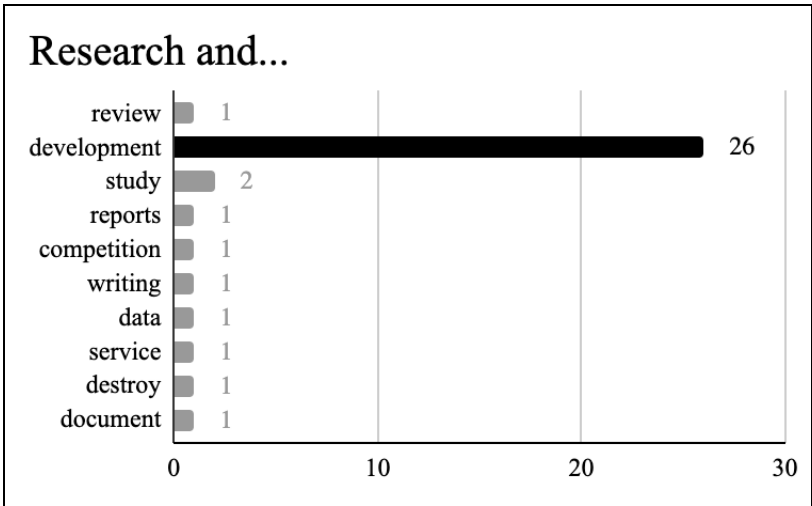


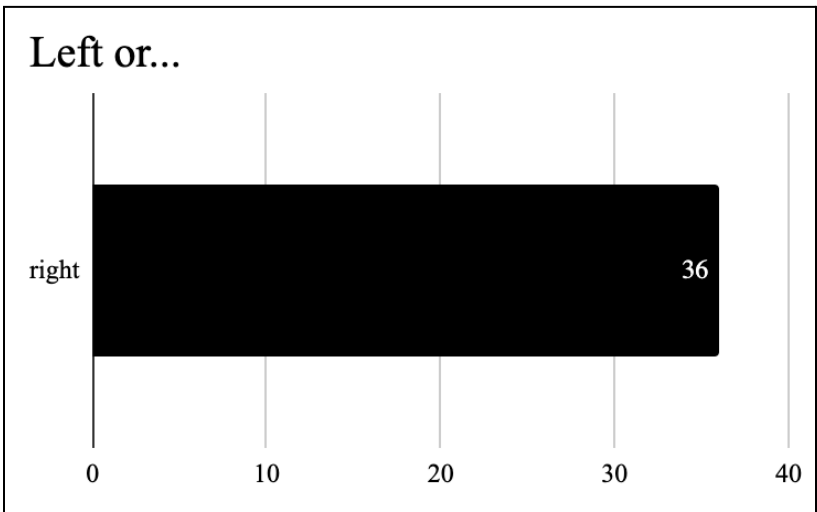
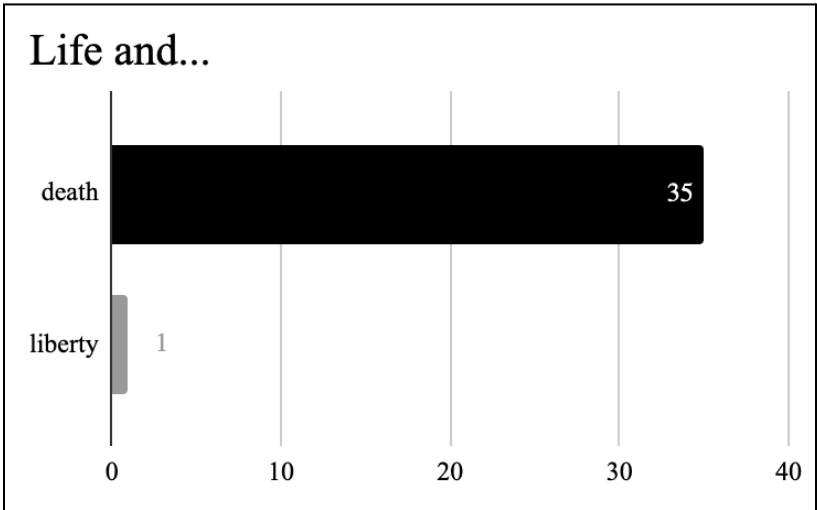
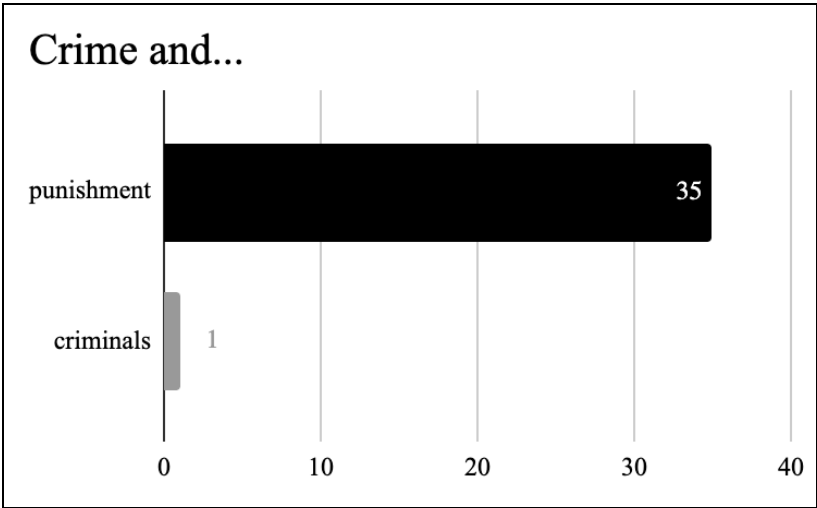


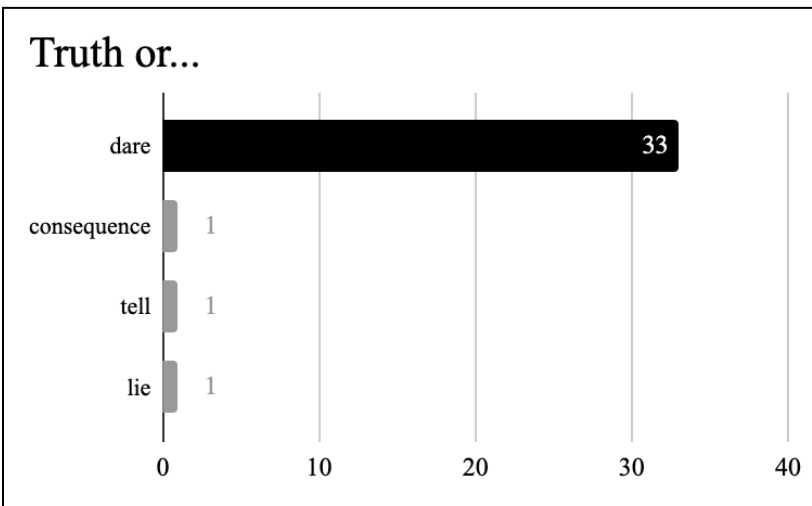
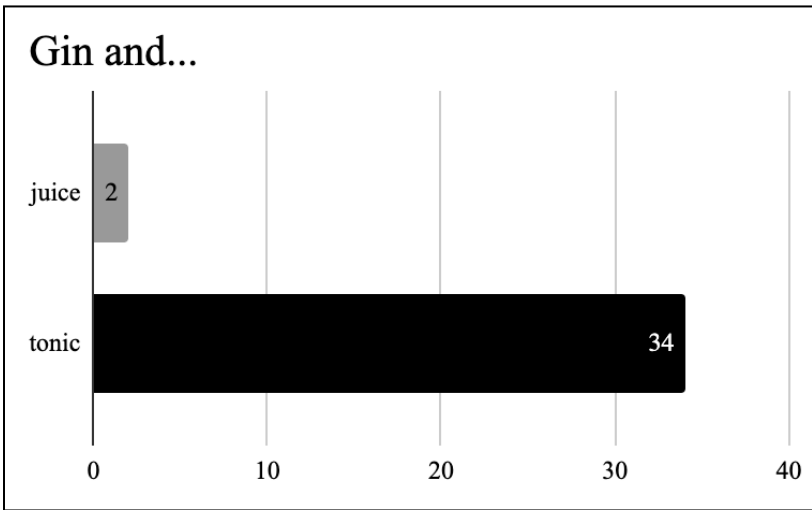
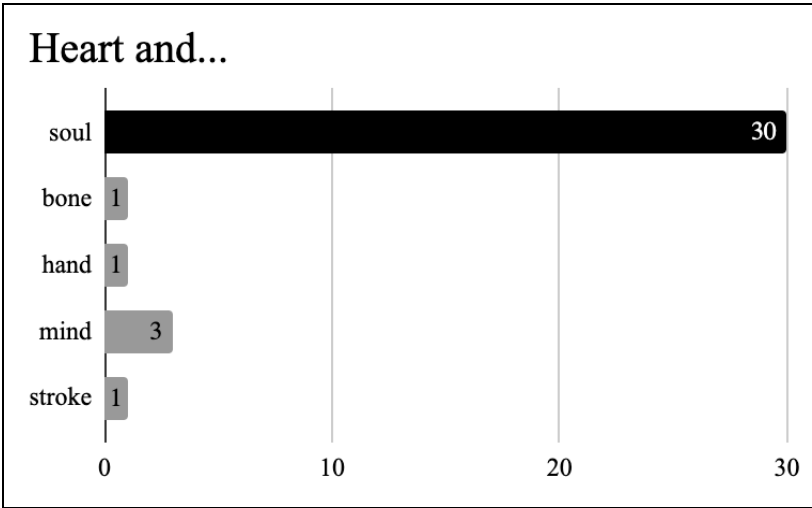


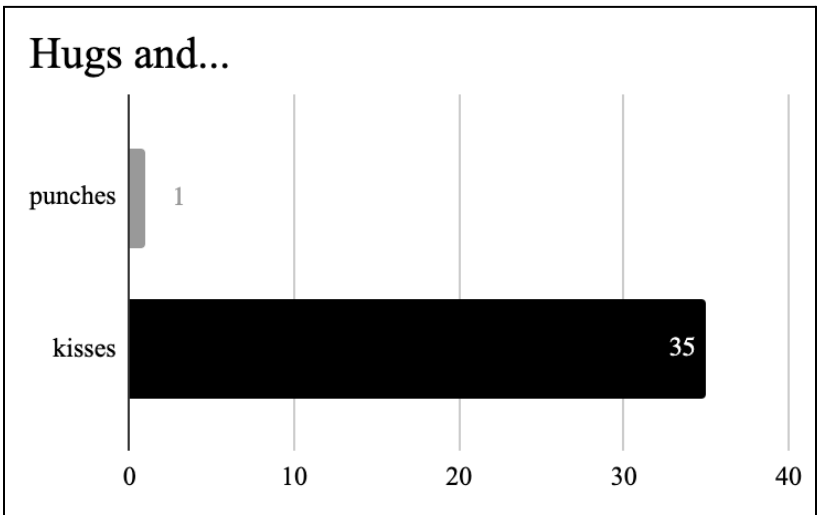
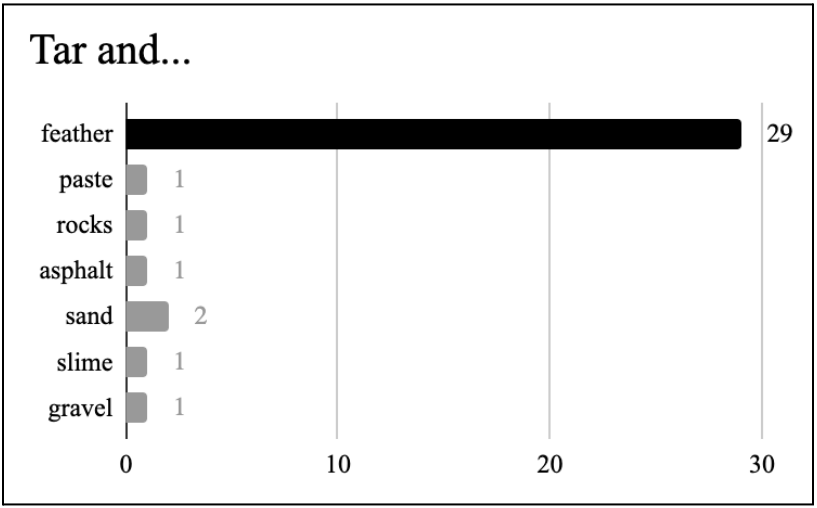
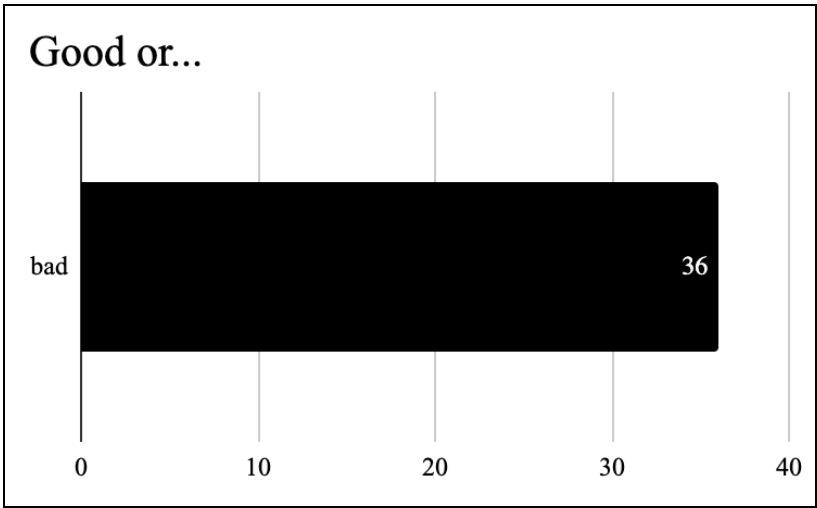












## Appendix D: Additional Transcription Error Tables

Table D.1: Transcription errors by condition

Condition	Autocorrect	New phrase	Nonword(s)	Wrong sound	Deletion
Control	0	0	2	1	0
Voicing	39	9	9	8	0
Place	28	10	10	17	2
Manner	13	14	11	6	2
Total	80	33	32	32	4

Table D.2: Manner transcription errors

Condition	Autocorrect	New phrase	Nonword(s)	Wrong sound	Deletion
Strident	3	9	4	2	2
Continuant	6	2	4	4	0
Nasal	2	1	2	0	0
Lateral	3	2	0	0	0
Total	14	14	10	6	2

Table D.3: Fricative transcription errors

Condition	Autocorrect	New phrase	Nonword(s)	Wrong sound	Deletion
Strident	2	9	4	2	2
Non-strident	5	0	2	2	0
Total	7	9	6	4	2

## References

- Alderete, J. & Davies, M. (2019). Investigating perceptual biases, data reliability, and data discovery in a methodology for collecting speech errors from audio recordings. *Language and Speech*, 62(2), 281-317.
- Alderete, J., & Tupper, P. (2018). Phonological regularity, perceptual biases, and the role of phonotactics in speech error analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 9(5), E1466.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.  
<https://doi.org/10.18637/jss.v067.i01>
- Benor, S. B., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 82(2), 233-278.
- Christensen, R. H. B. (2022). *ordinal – Regression Models for Ordinal Data* (Version 2022.11-16) [R package]. <https://CRAN.R-project.org/package=ordinal>
- Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *The Journal of the Acoustical Society of America*, 64(1), 44-56.
- Conklin, K., & Carrol, G. (2021). Words go together like ‘bread and butter’: The rapid, automatic acquisition of lexical patterns. *Applied Linguistics*, 42(3), 492-513.

- Cutler, A. (1981). The reliability of speech error data. *Linguistics*, 19(7-8), 561-582.
- Delaney-Busch, N., Morgan, E., Lau, E. & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, 187, 10-20.
- Eaton, C. T., & Newman, R. S. (2018). Heart and \_ or give and \_? An exploration of variables that influence binomial completion for individuals with and without aphasia. *American Journal of Speech-Language Pathology*, 27(2), 819-826.
- Ferber, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research*, 20, 105–122.
- Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, 30(2), 139-162.
- Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 47(1), 27-52.
- Lovitt, A., & Allen, J. B. (2006). 50 years late: Repeating Miller-Nicely 1955. In *Proceedings of the Ninth International Conference on Spoken Language Processing*, 2154-2157.
- Marin, S., Pouplier, M., & Harrington, J. (2010). Acoustic consequences of articulatory variability during productions of /t/ and /k/ and its implications for speech error research. *The Journal of the Acoustical Society of America*, 127(1), 445-461.



- Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, *189*(4198), 226-228.
- Martin, A. & Pepperkamp, S. (2015). Asymmetries in the exploitation of phonetic features for word recognition. *The Journal of the Acoustical Society of America*, *137*(4), EL307-EL313.
- Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338-352.
- Morgan, E. & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384-402.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. <http://w3.usf.edu/FreeAssociation/>.
- Potter, M. C., Moryadas, A., Abrams, I., & Noel, A. (1993). Word perception and misperception in context. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 3-22.
- R Core Team. (2023). *R: A language and environment for statistical computing*. Foundation for Statistical Computing. <https://www.R-project.org/>
- Steriade, D. (2001). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. *Ms., UCLA*.

Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). <https://doi.org/10.17605/OSF.IO/MD832>