Listening to Writers and Riders: Partial Contrast and the Perception of Canadian Raising

Patrick Joseph Murphy

A thesis submitted in conformity with the requirements

for the degree of Doctor of Philosophy

Graduate Department of Linguistics

University of Toronto

Abstract

Listening to Writers and Riders: Partial Contrast and the Perception of Canadian Raising

Patrick Joseph Murphy

Doctor of Philosophy

Graduate Department of Linguistics

University of Toronto

2019

Listeners generally have a greater perceptual sensitivity to native contrasts compared to allophones (Whalen et al., 1997; Boomershine et al., 2008) or non-native contrasts (Goto, 1971; Sundara et al., 2006) in discrimination and other tasks. Recent research has emphasized the gradient nature of contrast, showing that many phonological relationships are intermediate or variable between contrast and allophony (Hall, 2009, 2013). This dissertation presents a series of experiments investigating the perception of what has been called marginal contrast or partial contrast using Canadian Raising as a testing ground.

Experiment 1 tests discrimination of raised and non-raised diphthongs ([ʌj]~[aj] and [ʌw]~[aw]) in different phonological environments, finding better discrimination in the contrastive environment where they can create different words than in the allophonic environment where they cannot, but only for one of the two diphthongs (/aj/ but not /aw/). This diphthong difference was ambiguous—it could be a property of the diphthongs themselves, or it could have been a result of the stimuli used, specifically that [ʌj]~[aj] has more recognizable minimal pairs (e.g., *writing/riding*) than [ʌw]~[aw] (e.g., *clouting/clouding*). Experiments 2, 3, and 4 clarify this partial contrast effect and diphthong difference, finding support for an inherent diphthong difference (using non-words in Experiment 2) and for an additional effect of the minimal pairs (Experiments 3 and 4). Experiments 1b, 1c, 2b, 3b, and 4b are semi-replications of these initial four experiments. They lack an additional experimental condition that was present in the original experiments, and in each case the original partial contrast effect fails to replicate, suggesting that partial contrast effects depend on quality/quantity of linguistic exposure. Finally, Experiment 5 tests discrimination of Canadian Raising diphthongs by Canadians and Americans, finding generally faster and more accurate discrimination by Canadians, with differences between different American regions as well.

Together, these experiments provide insight first and foremost into the effect of contrast—specifically partial contrast—on discrimination, as well as other topics such as cross-dialectal perception (and the effect of dialect stereotypes and dialect exposure on perception) and regional differences in the production of raising (and related phenomena) in Canada and the United States.

Acknowledgments

Graduate school has been an incredible experience for me, and many people deserve credit for that. The obvious person to start with is Phil Monahan, who had more influence than anyone else on my development as a researcher. He was my dissertation supervisor, a co-supervisor on both of my generals papers, and he taught the class on speech perception that really inspired the research path I took. I've appreciated his intellectually rigorous but down-to-earth supervising style, his ability to creatively brainstorm and sort through ideas with me, and his support for my first time as a course instructor (LINB29 Quantitative Methods in Linguistics at UofT Scarborough), which was an extremely rewarding experience.

The other members of my dissertation committee also have my enthusiastic appreciation. Jessamyn Schertz has a remarkable skill of making unexpected or confusing results make sense (as I saw with my first generals paper and then this dissertation), and I must highlight that she and her advanced quantitative methods class were central in developing my skills and confidence working with data and statistics. Next, Jack Chambers has an impressive eye for detail, a wealth of knowledge on North American dialects, and an enviable level of energy and inquisitiveness even in retirement. He also pioneered work on Canadian Raising that made this dissertation possible in the first place. Finally, my internal and external reviewers (Yoonjung Kang, Nathan Sanders, and Kathleen Currie Hall) provided constructive comments and insightful perspectives that greatly benefited this dissertation and my thinking about these topics.

Going back, I want to mention my master's supervisor, Diane Massam, who was a big part of why I enjoyed my MA so much and decided to continue on to a PhD. I also want to give a shout-out to my 2013/2014 UofT linguistics MA cohort. They've since spread to a lot of different places, but there was something especially exciting about that first year of graduate school that we all experienced. To go back even further, I'm also indebted to the linguistics faculty at Saint Mary's University in Halifax (Daniel Currie Hall, Elissa Asp, and Egor Tsedryk) for an undergraduate experience that I remember very fondly as well.

I'm also thankful to many special people outside of academia. My parents, Michele and Darrin, provided me with one of the best childhoods anyone could hope for, which I have to assume plays a role in anything I achieve. My grandmother, Sylvia, always supported me and provided perspective by reminding me that getting a PhD is actually a pretty cool thing to do, which wasn't always at the front of my mind as I was deep into the details of a research project. Finally, to Jasmine, thank you for being my companion for these amazing years in Toronto, for supporting me through the ups and downs, and for pushing me to get out, explore, and develop as a person instead of just as an academic.

Contents

List of Tables

List of Figures

Chapter 1

Introduction & Background

This dissertation is a psycholinguistic investigation of intermediate phonological relationships and their impact on speech perception. An intermediate phonological relationship, also known as marginal contrast, partial allophony, or partial contrast (the term that will be used in this dissertation), occurs when the phonological relationship between two sounds appears to be intermediate or variable between contrast and allophony (Hume and Johnson, 2003; Ladd, 2006; Hall, 2013). Various examples of partial contrast can be found across languages, but the phenomenon of interest for this dissertation is Canadian Raising, a dialectal feature characteristic of Canadian English but also present to some extent in the United States (Chambers, 1973; Niedzielski, 1999; Labov et al., 2006). Although canonically an allophonic alternation, with raised diphthongs [ʌj, ʌw] occurring before voiceless consonants (e.g., [ˈɹʌjt] *write*) and non-raised diphthongs [aj, aw] occurring elsewhere (e.g., [ˈɹajd] *ride*), the interaction between Canadian Raising and North American /t/–/d/ flapping (Vaux, 2000) results in the possibility of [ʌj]~[aj] and [ʌw]~[aw] minimal pairs in the context of a following flap, such as [ˈɹʌjɾɪŋ] *writing* (underlying /t/) and [ˈɹajɾɪŋ] *riding* (underlying /d/). This dissertation will present a series of five psycholinguistic experiments (with an additional five semi-replications of those experiments) testing how the discrimination of raised and non-raised diphthong variants ([ʌj]~[aj] and [ʌw]~[aw]) is influenced by the local phonetic context—especially whether the following consonant is a flap—and the resulting change in their phonological relationship, as raised and non-raised diphthong variants have more of a contrastive status when followed by a flap, compared to their regular status as allophones elsewhere. This study is grounded in the extensive previous literature on the perceptual impact of contrast as compared to allophones and non-native contrasts (e.g., Liberman et al., 1957; Lisker and Abramson, 1970; Goto, 1971; MacKain et al., 1981; Boomershine et al., 2008) and the more recent developing body of perception literature that has focused on partial contrast specifically (Hume and Johnson, 2003; Celata, 2008; Murphy et al., 2016; Stevenson and Zamuner, 2017).

1.1 Contrast and Speech Perception

The critical early stage of speech perception receives incoming speech as a continuous acoustic waveform and extracts from it the discrete phonological units—consonants like /p/ and /b/ and vowels like /ɑ/ and /i/—that are widely believed to be the basic building blocks of words in our long-term memory or mental lexicon (Studdert-Kennedy, 1976; Phillips, 2001). While this process of extracting sound categories from the acoustic signal is a universal component of human speech perception, the categories themselves vary between languages. Language-specific systems of sound categorization are studied in phonology under the concept of contrast, "the opposition between distinctive sounds in a language" (Dresher, 2011). For example, some languages have a three-way voicing contrast in stops, distinguishing between voiced, voiceless, and aspirated stops (like Thai and Vietnamese); other languages have a two-way distinction between voiceless and aspirated stops (English and some dialects of German) or between voiced and voiceless stops (Russian and Turkish) (Cho et al., 2019). Some languages, such as a large number of Australian Aboriginal languages, do not make a voicing distinction in stops at all (Austin, 1988). However, the fact that a language lacks a sound as a distinct phonemic category does not necessarily mean that the sound does not occur in the language. For example, the glottal stop [ʔ] is commonly used in English (as in [ˈbʌʔn̩] *button*, although the exact environments where it can be used depend on the dialect) but [ʔ] is not a sound category in its own right; instead, [ʔ] is an allophone of the phoneme /t/, or a variant of the sound category /t/ (Roberts, 2006; Eddington and Taylor, 2009; Seyfarth and Garellek, 2015). The distinction between [ʔ] and another variant of /t/, like aspirated [t^h], can never be used to distinguish between words.

These language-specific systems of contrast, including which sounds are contrastive (phonemes in their own right) and which sounds are not contrastive (which can include allophones of the same phoneme, or a contrast from another language that does not exist in the listener's language), have been found to have behavioural consequences both for the identification and discrimination of sounds. These effects fall within the broader phenomenon of categorical perception, where an observer's mental categories influence their perception (Goldstone and Hendrickson, 2010). Categorical perception effects have also been studied in the auditory domain for musical patterns and properties (Burns and Ward, 1974; Fiske, 1997) and in the visual domain for colour perception (Bornstein and Korda, 1984; Winawer et al., 2007;

Roberson et al., 2008) and perception of facial expressions (Etcoff and Magee, 1992; Young et al., 1997; Fugate, 2013).

The effect of contrast on identification has frequently been studied by creating continua that start at one sound and, with each additional step, increasingly become closer to a target sound. The task of participants is to identify which sound they hear at each step. The effect of contrast in these studies is that boundaries between contrastive sounds (in the listener's language) are associated with sharp (rather than gradual) shifts in perception from one category to another. For example, Liberman et al. (1957) created a 14-step continuum from [be] to [de] to [ge] by varying the second formant transition, which plays a large role in the perception of place of articulation of stop consonants. They played these 14 syllable variations to English speakers and asked them to identify the consonant they heard among ⟨b, d, g⟩, finding that the participants had a tendency to divide the continuum into "three sharply bounded phonemic categories": the first third of the continuum (around steps 1 to 3) was perceived fairly consistently as /be/, the middle third (around steps 4 to 9) as /de/, and the last third (around steps 10 to 14) as /ge/. Similarly, Lisker and Abramson (1970) found generally sharp perceptual boundaries between categories on a stop VOT (-150 ms to +150 ms) continuum for Spanish, English, and Thai listeners, although the location of the boundaries differs by language. However, the actual effect of contrast is most clear when comparing such results to listeners who do not have the category boundary (i.e., who do not have the two ends of the continuum as being contrastive in their language), as seen in the results of MacKain et al. (1981) in Figure 1.1 with English and Japanese speakers. They created a 10-step continuum from [ɹ] (in *rock*) at step 1 to [l] (in *lock*) at step 10 by varying spectral characteristics of the second and third formant, and temporal characteristics of the first formant. As seen by the percentage of [ɹ] responses plotted, the English speakers have a sharp shift in perception (from [ɹ] to [l]) in the middle of the continuum, with relatively few points in the continuum that are ambiguous and inconsistently labelled. On the other hand, Japanese speakers, who do not have a categorical distinction between [ɹ] and [l] in their language,[1] have a much more gradual shift in perception, with a greater degree of inconsistency in the middle (and even at the ends of the continuum).

As for the effect of contrast on discrimination, it has been found that listeners more accurately and more quickly discriminate between sounds that come from different categories or phonemes (in their language) than between sounds that would fall under the same phoneme, which includes phonetic and allophonic variants of phonemes in their language as well as non-native phonemes that most closely map onto the

[1] Japanese has one /r/ phoneme whose realization depends on factors that include phonological environment and dialect. Documented realizations include the alveolar rhotic approximant [ɹ], short-duration alveolar lateral approximant [l̆], alveolar flap [ɾ], and alveolar lateral flap [ɺ] (Magnuson, 1998).

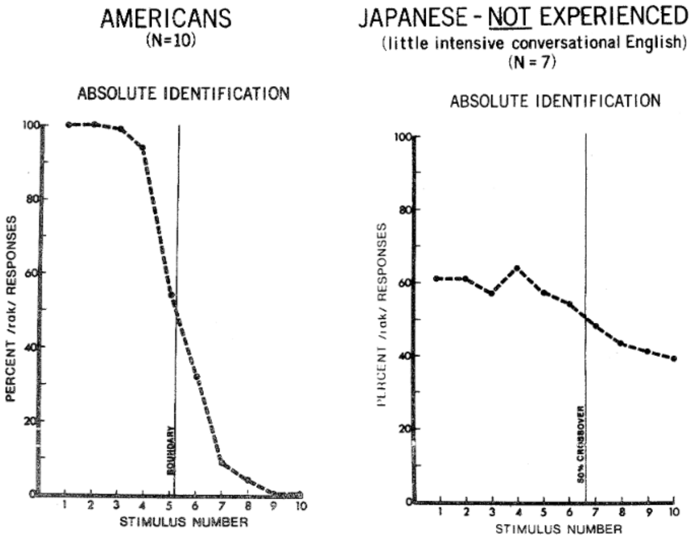

Figure 1.1: Identification of an English [ɹ] to [l] continuum by English speakers (L) and Japanese speakers (R) from MacKain et al. (1981) (their Figures 2 and 3)

same native phoneme. Liberman et al. (1957) used the same 14-step continuum (from [be] to [de] to [ge]) and tested participants on their ability to discriminate between adjacent or nearly adjacent steps (e.g., discriminating between steps 1 and 3 or between steps 3 and 5). They used an ABX discrimination task, which involves participants hearing the two sounds from the continuum (A and B) followed by a repetition of one of the previous sounds (X) that they have to identify as being either the first or the second sound (Munson and Gardner, 1950). In this experiment, participants were better able to discriminate between steps that fell across category boundaries (e.g., between steps 3 and 5) than steps that fell within the same category (e.g., 1 and 3). Crucially, this discrimination pattern would not be predictable based purely on the acoustics of the stimuli (the physical difference between steps 1 and 3 is the same as between steps 3 and 5), but it is predictable from the category boundaries indicated in the previous experiment.

Similar findings are visible in the results of Lago et al. (2015), who worked with an 11-step English fricative continuum from [ʃ] to [s], created by manipulating the location of the fifth and sixth formants. As seen in the left panel of Figure 1.2, which shows how often each step on the continuum was identified as [s], the continuum was overwhelmingly identified as [ʃ] for steps 1 to 6 and identified as [s] for steps 8 to 11, marking the category boundary at step 7. The right panel shows the result of an AX (or same-different) discrimination task, in which participants heard two sounds (two steps on the continuum, or one step on the continuum repeated twice) and had to determine whether they were the same or different.

Discrimination accuracy on the y-axis is plotted as *d′* (the sensitivity index), a measure of performance that takes into account the possibility of participants being biased towards responding “same” or “different” (Macmillan and Creelman, 2004). This plot shows the accuracy of discriminating sounds on the continuum that are 2, 4, 6, or 8 steps apart. One pattern observed is better discrimination for steps that are further apart, which is expected from the fact that there is a greater physical difference between steps that are further apart. Another pattern in these results is that discrimination is more accurate between steps that cross (or approach) the category boundary at step 7 (which was determined in the identification task) than for steps that are clearly in one category or the other. This is particularly visibile in the two-step discrimination results, which find participants being much more capable of noticing the difference between steps 6 and 8 than between steps 1 and 3.

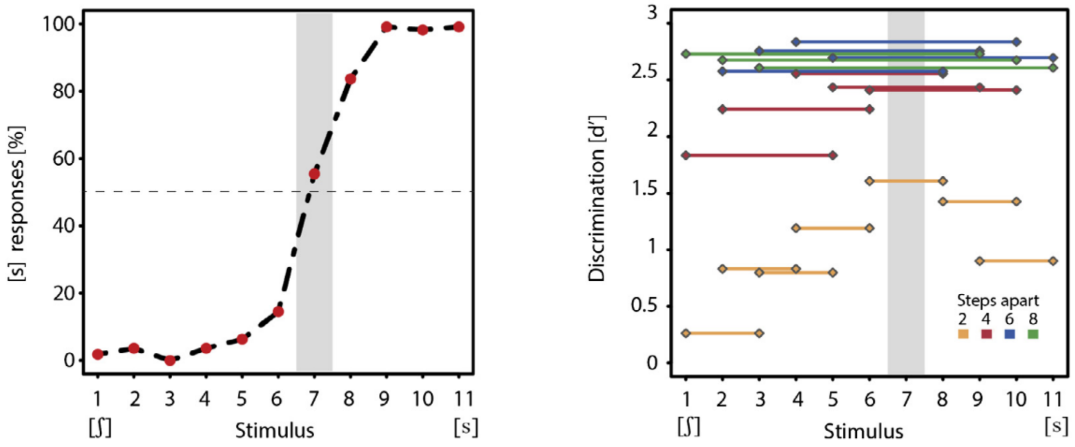

Figure 1.2: Identification (L) and discrimination (R) of a fricative continuum from Lago et al. (2015) (their Figure 1)

These studies used a continuum between two phonemes in the listeners’ language and found better discrimination for pairs of sounds on that continuum that crossed phoneme boundaries than pairs that fell within the same phoneme boundary. Other studies have demonstrated better discrimination for native phonemes over native allophones, as well as the difficulty of discriminating non-native contrasts that would be mapped onto the same native phoneme, using more natural tokens that were not modified to be on a continuum.

Whalen et al. (1997) tested English speakers on discrimination of [p^{h}], [p], and [b] word medially, finding that accuracy was lower for [p^{h}]~[p] (which are allophones of /p/)[2] than for the other two pairs. These results (from two AXB discrimination experiments) are presented in Figure 1.3, with percentage correct

[2] The aspirated allophone occurs in medial position before stressed vowels and the unaspirated allophone occurs medially before unstressed vowels.

on the y-axis. Boomershine et al. (2008) found that Spanish listeners more quickly discriminate [d]~[ɾ] (contrastive in Spanish, allophonic in English) while English listeners more quickly discriminate [d]~[ð] (contrastive in English, allophonic in Spanish). Similar findings with [d]~[ɾ] and [d]~[ð] are replicated in discrimination accuracy by Barrios et al. (2016) and also with Spanish- and English-learning infants by Meredith and Maye (2009). Boomershine et al. (2008) also tested ratings of similarity, finding that Spanish listeners rated [d]~[ɾ] (contrastive in Spanish) as more different than English speakers did, while English speakers rated [d]~[ð] (contrastive in English) as more different than Spanish speakers did. Johnson and Babel (2010) tested English and Dutch speakers on discrimination of [s]~[ʃ] (contrastive in English, allophonic in Dutch) and actually did not find a difference between the two language groups in discrimination, but they used a speeded discrimination task with a short (100 ms) interstimulus interval (ISI) with the intention of recruiting low-level auditory perception instead of a more phonetic or phonemic kind of processing that is more likely to be influenced by their language's phonology (Werker and Logan, 1985). In their separate similarity rating experiment, Johnson and Babel (2010) found that Dutch speakers rated [s]~[ʃ] as being more similar than English speakers did, which potentially indicates that there would be a group difference in discrimination under different experimental conditions.

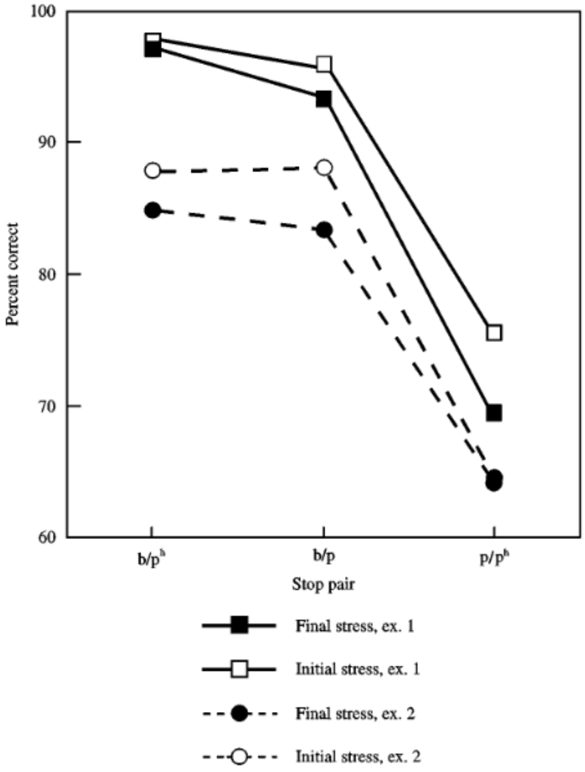

Figure 1.3: Discrimination of stop pairs in English in two experiments from Whalen et al. 1997 (their Figure 2)

As for the difficulty of discriminating non-native contrasts, there is the difficulty faced by Japanese speakers on discriminating the English /ɹ/~/l/ contrast (Goto, 1971), English speakers on the Hindi /t/~/ʈ/ and /tʰ/~/dʰ/ contrasts (Werker et al., 1981), French speakers on the English /d/~/ð/ contrast (Sundara et al., 2006), and English speakers on the Polish /ʂ/~/ɕ/ contrast (McGuire, 2007). However, discrimination of non-native contrasts is better if the non-native sounds are mapped onto different native phonemes (Best et al., 2001) or perceived as non-speech, as has been found for perception of clicks (Best et al., 1988, 2003). The different ways that non-native sounds can map onto native phonemes are described in the Perceptual Assimilation Model (Best et al., 2001). Developmentally, this difficulty discriminating non-native contrasts (compared to native contrasts) emerges between the ages of 6–12 months (Tsushima et al., 1994; Best and McRoberts, 2003).

It should also be noted that not all speech sounds produce effects associated with contrast or categorical perception equally, in identification or discrimination (Fry et al., 1962; Pisoni, 1975; Kröger et al., 2011; Chan et al., 1975; Abramson, 1977; Francis et al., 2003). Consonants tend to be perceived most categorically while vowels and tones tend to exhibit weaker categorical perception effects. Explanations in the literature have linked this to better auditory short-term memory for the acoustic properties of vowels (Fujisaki and Kawashima, 1970; Pisoni, 1973) and more variation in the acoustic realizations of vowels than of consonants; for example, vowel formants vary significantly due to vocal tract differences while stop voice onset time does not (Huang, 2010; Kronrod et al., 2016).

1.2 The Role of Context in Speech Perception

There is considerable evidence that listeners make use of local phonetic context in speech perception. One particularly well-studied example is the effect of vowel formant transitions in consonant perception. Consonants can affect the formant frequency at the beginning of the following vowel (or the end of a preceding vowel), and it has been found that listeners use this information in the vowel signal for identification of consonants.[3] For example, formant transitions play a role in distinguishing the place of articulation of oral stops in English (Liberman et al., 1954; Harris et al., 1958; Menon et al., 1974), in Hebrew (Kishon-Rabin et al., 2011), in Danish (Fischer-Jorgensen, 1972), and in Dutch (Smits et al., 1996). Formant transitions have also been found to influence perception of place of articulation for nasal stops (Liberman et al., 1954)

[3] Formant transitions are important enough for the perception of certain consonants that many of the continuum examples from Section 1.1 were created by modifying formant transitions.

and fricatives (Harris, 1958; Mann and Repp, 1980; Repp, 1981) in English. Languages vary in terms of acoustic cues used; Polish, for example, appears to rely less on formant transitions in the perception of stop consonants than other languages (Aperlinski and Schwartz, 2015). Other contextual cues for consonants include preceding vowel duration affecting perception of consonant voicing in English (Raphael, 1972; Luce and Charles-Luce, 1985) and, to a lesser extent, in Dutch (Broersma, 2010), preceding vowel nasality affecting perception of consonant nasality in English (Lahiri and Marslen-Wilson, 1991), as well as the laryngeal contrast (lenis, fortis, aspirated) in Korean being cued by the fundamental frequency of the onset of the following vowel (Han and Weitzman, 1970; Cho et al., 2002; Lee and Katz, 2016; Schertz et al., 2019). In addition to the perception of consonants being influenced by neighbouring vowels, it has also been found that the perception of vowels is influenced by the neighbouring consonant context, such as the perception of English vowels being influenced by the direction and rate of formant movements in adjacent glide consonants (Lindblom and Studdert-Kennedy, 1967; Williams, 1987) and by properties of spectral energy in the burst of an adjacent stop consonant (Nearey, 1989; Holt et al., 2000).

1.3 Partial Contrast

1.3.1 Defining Partial Contrast

As explained in Section 1.1 on contrast, sounds can exist in a language as phonemes in their own right or as allophones of another phoneme. Hall (2013) provides a detailed documentation of the criteria traditionally used to determine whether two sounds are separate phonemes or allophones of the same phoneme. The first criterion, predictability of distribution, says that two sounds are contrastive (different phonemes) if, for at least one phonological environment in the language, it is impossible to predict which of the two sounds will occur. If this is not the case, and they can be predicted in all of the environments where they occur, the two sounds are allophonic. For example, in English /ɹ/ and /l/ can both occur in the same phonological environments (such as onsets and codas), making them unpredictable and thus different phonemes. However, clear [l] and dark (velarized) [ɫ] both occur in many English dialects but they are at least traditionally described as following the pattern of [l] in onsets and [ɫ] in codas (Sproat and Fujimura, 1993). The predictable pattern of their occurrence defines them as allophones.

The second criterion that Hall (2013) presents for determining whether two sounds are separate phonemes or allophones of the same phoneme is lexical distinction, or the commutation test, which says that two sounds are contrastive if the substitution of one for the other in some phonological environment creates a different word. For example, switching from /l/ to /ɹ/ in the English word *light* creates a different word (*right*), showing /l/ and /ɹ/ to be different phonemes. However, switching the initial sound in *light* from a clear [l] to a dark [ɫ] might result in an unexpected pronunciation (depending on the dialect) but it does not create a different word, and there are no examples where it does, showing those two sounds to be allophones. Hall (2013) outlines several additional criteria for determining whether two sounds are separate phonemes or allophones of the same phoneme (such as the judgement by native speakers of whether they are "the same sound"), but she describes these initial two (predictability of distribution and lexical distinction) as the most important.

One difficulty of these criteria, as she explains, is that they sometimes conflict with each other. For example, in Scottish English, the [ɑj]~[ʌj] distinction (not to be confused with Canadian Raising, introduced below) is phonetically the only difference between *tied* [ˈtɑjd] and *tide* [ˈtʌjd], but their distribution is predictable: [ɑj] morpheme-finally and [ʌj] morpheme-internally. Going by the lexical distinction test these sounds are contrastive, but they are allophonic according to the predictability test. Even a test on its own does not always produce a clear result. In the lexical distinction test, finding a minimal pair is considered strong evidence for contrastiveness, while failing to find a minimal pair is weaker evidence for allophony, due to lexical gaps (which mean that phonologists often rely on near minimal pairs).

An additional problem that she raises is that these tests classify all relationships as being either contrastive or allophonic, even though many phonological relationships have elements associated with both contrast and allophony. Such intermediate relationships have been referred to as partial contrast, marginal contrast, or partial allophony, among other terms. One example from Hall (2013) is [r] and [ɾ] in Spanish, which have minimal pairs intervocalically (e.g., *carro* "car" and *caro* "expensive") but are otherwise predictably distributed: only [r] occurs word-initially and after a heterosyllabic (i.e., different syllable) consonant, while only [ɾ] occurs after a tautosyllabic (i.e., same syllable) consonant and word-finally before a vowel (in the following word) (Hualde, 2004).[4] Under the traditional binary distinction, this relationship would be considered contrastive because of the intervocalic environment where the two sounds are unpredictable and minimal pairs are possible. However, this case is clearly different from contrast in the canonical sense—full contrast—due to the other environments where [r] and [ɾ] are predictably distributed. Another such example is [x] and [ç] in German, which are normally predictably distributed. However, there are a limited number of minimal pairs (e.g., *kuchen* [ˈku.xən] "cake" versus *kuhchen* [ˈku.çən] "little cow") based

[4] In two other environments, before a consonant and before a pause, Hualde (2004) describes the rhotic as "indistinct/variable".

not on phonological environment but instead on morphology (*kuhchen* “little cow” has the diminutive *-chen* suffix) (Hall, 2013).

Partial contrast should probably be understood along a continuum from full contrast to full allophony, rather than a third category (Hall, 2009, 2013). Some partially contrastive relationships fall closer to full allophony, like the examples above, which are primarily allophonic but with a contrast that appears under certain limited conditions. Other partially contrastive examples fall closer to full contrast, such as when a normally contrastive sound pair has its contrast neutralized under certain conditions. For example, /t/ and /d/ are separate phonemes in English, but North American varieties neutralize[5] this contrast by turning both sounds to the alveolar flap intervocalically before unstressed vowels (although some sonorants can also precede or follow; see Vaux, 2000). Mid vowels in Romance provide another example of partial contrast where a contrast is neutralized. In Italian, higher mid /e, o/ and lower mid /ɛ, ɔ/ vowels are neutralized to higher mid vowels in unstressed syllables (Ladd, 2006).

Sometimes partial contrast can arise due to loan words. For example, Japanese does not normally have a contrast between [t] and [cɕ] before a high front /i/ (palatalization would turn /ti/ to [cɕi]). However, certain loan words are “unassimilated” and exempt from this process (e.g., [tiːn] “teen(ager)”) (Hall, 2013).

1.3.2 Partial Contrast and Perception

While the behavioural consequences of language-specific sound categories or phonemes (and, to a lesser extent, language-specific allophonic alternations) have received significant attention in the literature (Section 1.1), the perception of partial contrast is a newer and less-studied topic. The existing research will be reviewed here, first from two studies in identification and then two studies in discrimination.

The first study on identification that will be covered involves Laurentian French (Quebec French and historically related varieties), where stops allophonically become affricates (/t, d/ ⟶ [ts, dz]) before high front vowels (/i, y/); however, there exist a small number of lexicalized or non-allophonic affricates without such a place restriction ([dzaʁ] *tsar* “tsar”, [tse] *tsé* “y’know”). Murphy et al. (2016) found that speakers of Laurentian French perceive a 10-step affricate to stop ([ts] to [t] and, separately, [dz] to [d]) continuum as having a sharper perceptual boundary in the non-allophonic context (before an [e]) than in the allophonic context (before an [i]), as seen in Figure 1.4. In the contrastive environment (compared to the allophonic

[5] This has been found to be incomplete neutralization. Vowels preceding /t/⟶[ɾ] flaps are slightly shorter in duration (~9 ms) than vowels precedeing /d/⟶[ɾ] flaps (Braver, 2013), although Braver (2014) finds this difference to be imperceptible.

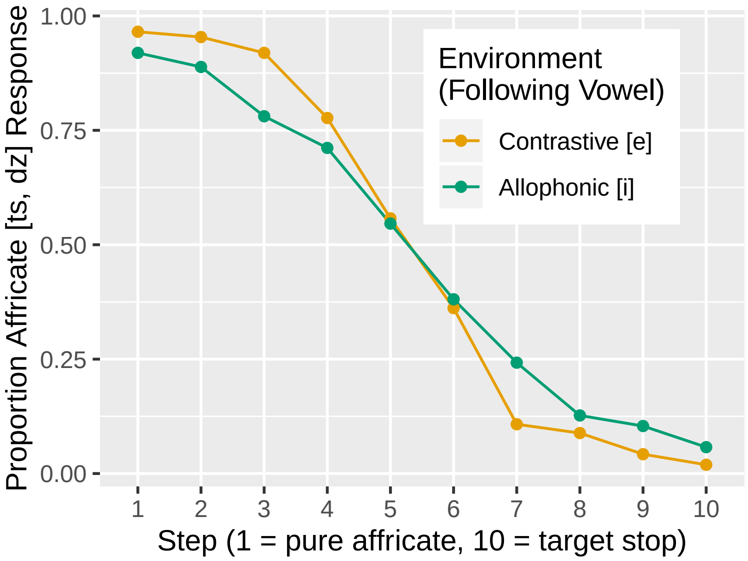

Figure 1.4: Identification of an Affricate-Stop Continuum (Proportion Affricate Response) from Murphy et al. (2016)

environment), the more affricate-like side of the continuum (steps 1–4) produces more consistent affricate responses while the more stop-like side of the continuum (steps 6–10) produces more consistent stop responses.

The second study in identification involves the Western Tuscan dialect of Italian, which has /ts/ and /s/ neutralized to [ts] after sonorants. Celata (2008) found that speakers of this Italian dialect are slower, less accurate, and less confident when asked to identify /ts/ and /s/ in the neutralized environment (after a sonorant) than in other environments, both in Italian non-words and Russian real words (functionally non-words given that the participants did not have experience with Russian), while speakers of Northern Italian (who do not have this neutralization in their dialect) did not exhibit any such environment differences. The Italian non-words were tested on a forward gating paradigm, where participants hear a small portion of the beginning of the word, then increasingly larger framents of the word (still starting at the beginning) until they hear the whole word, giving their identification response (and a level of confidence) for each fragment. The Russian words were presented in a simpler identification task, although this task still involved participants indicating their level of confidence in their judgement.

The first study on discrimination that will be covered involves pairs of vowels in Laurentian French that, based on a corpus analysis of predictability of distribution, have a high level of contrast ([a]~[ɔ]), a medium level of contrast ([o]~[ʊ]), or a low level of contrast ([y]~[ɤ]). Stevenson and Zamuner (2017) measured accuracy and reaction time on an AX (same-different) vowel discrimination task finding overall faster and more accurate discrimination for the high contrast vowel pair, followed by the medium contrast vowel pair, and then the low contrast vowel pair (although the effect was clearer and more consistent for accuracy). Under the traditional binary view of contrast, the high and medium contrast sound pairs would be classified as contrastive while the low contrast sound pair would be classified as allophonic. However, this study found behavioural evidence for the medium contrast sound pair being distinct from both of the others.

The second discrimination study involves the Mandarin process of third tone sandi, which results in the contrast between the second tone (also called 35, which means that it rises from 3 to 5 on a five-point pitch scale) and third tone (also called 214) being neutralized when followed by another third tone. Hume and Johnson (2003) measured reaction time on an AX tone discrimination task and find that Mandarin speakers are slower to discriminate between the second and third tone than other tones, a finding that is only partially seen from the control group of American English speakers, suggesting that the finding for Mandarin speakers is not purely a result of the second and third tone being more acoustically similar than the other tones. The Mandarin speakers also exhibited a context effect, where the discrimination was slowest in the neutralizing environment, although it was no stronger than the context effect found for the American control group, indicating that this context effect can be explained by acoustic factors rather than partial contrast and the phonological status of the neutralizing environment. Figure 1.5 shows the perceptual distance between the four tones of Mandarin, with the relevant second and third tones on the left. The perceptual distance was calculated as 1 / reaction time, which means that a smaller distance between tones indicates a slower discrimination response. The thick black lines indicate the perceptual distances of the Mandarin speakers, and the thin black lines correspond to the American control group. The black-filled points correspond to the neutralizing environment, while the white-filled points are non-neutralizing environments.

Together, these four studies present two types of partial contrast effects on perception: what could be called general partial contrast effects, and context-based partial contrast effects. General partial contrast effects are findings in Stevenson and Zamuner (2017) and Hume and Johnson (2003) that involve participants responding differently to partially contrastive sound pairs than to fully contrastive or fully allophonic sound pairs. In Stevenson and Zamuner (2017), faster and more accurate discrimination was found for Laurentian French vowel pairs that were more contrastive compared to vowel pairs that were more allophonic or predictable. In Hume and Johnson (2003), faster discrimination was found for fully contrastive Mandarin tones than partially contrastive (neutralizing) tones. Context-based partial contrast effects, on

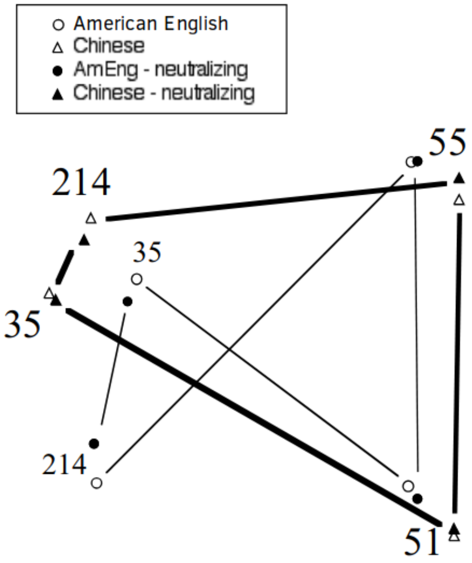

Figure 1.5: Perceptual Distance between Mandarin Tones from Hume and Johnson (2003) (their Figure 2)

the other hand, are findings in Murphy et al. (2016) and Celata (2008) that participants respond differently to partially contrastive sound pairs in contrastive environments than in non-contrastive (allophonic or neutralizing) environments. In Murphy et al. (2016), participants perceived a continuum between two consonant classes as having a sharper perceptual boundary in the contrastive environment than the allophonic environment. In Celata (2008), which is more comparable to the two discrimination tasks (because it tested speed and accuracy rather than perceptual preferences on ambiguous sounds on a continuum), participants from the dialect region that has the neutralization exhibited faster and more accurate identification of partially contrastive consonants in the contrastive environment than in the neutralizing environment. There also appeared to be a context-based partial contrast effect in Hume and Johnson (2003), because speakers of Mandarin performed more slowly at discriminating the partially contrastive tones in the neutralizing environment, but this finding was also found to a similar extent in the American English control group. Given that the American English speakers did not have knowledge of the phonology of Mandarin, this context effect is likely not related to partial contrast or any other phonological factor.

If the behavioural correlates of contrast in the traditional binary sense (contrastive versus non-contrastive), such as speed and accuracy of discrimination or identification, or the shape of identification curves on a continuum, are also affected by partial contrast, then the implication is that partial contrast is real and relevant in the sound systems of languages, much like contrast in the traditional binary sense. Thus, in finding these partial contrast effects in perception, these four studies provide behavioural evidence for the importance of understanding contrast as gradient (with a range of possible phonological relationships in between "contrastive" and "non-contrastive") rather than, or in addition to (Hall and Hall, 2016), seeing it as binary.[6] Partial contrast effects on perception have also been used as support for episodic or exemplar models of the lexicon (Stevenson and Zamuner, 2017), which reject the traditional assumption that storage of words in long-term memory takes the form of abstract lexical representations with the "noise" (speaker- or utterance-specific details) filtered out; these models instead propose that a significant amount of detail from linguistic experience is saved in long-term memory and is later used to mediate perception (Goldinger, 1996, 1998; Goldinger et al., 1999). These models more clearly accommodate a gradient understanding of contrast and categories (and the gradient behavioural results outlined above) because the existence of categories is driven by the frequency of encountered tokens that are mapped to that category, which means that categories can be more or less robust. These models can come in stronger or weaker forms, with weaker forms, which also allow a role for abstract lexical representations alongside episodic detail, probably being more tenable (McQueen et al., 2006; Monahan, 2009).

1.4 Canadian Raising

1.4.1 Phonetics and Geographic Distribution

Canadian Raising (Ahrend, 1934; Joos, 1942; Chambers, 1973) is a phenomenon whereby the diphthongs /aj/ and /aw/ start with a higher onset before a tautosyllabic voiceless consonant, as seen in 1–2.

1. *Ice* [ˈʌjs] (raised) compared to *eyes* [ˈajz]
2. *House* [ˈhʌws] (raised) compared to *houses* [ˈhawzɪz]

Labov et al. (2006) phonetically define Canadian Raising as a difference of at least 60 Hz in mean F1 values (i.e., vowel height) in the onset of the diphthong between voiceless and voiced environments. The phenomenon is characteristic of Canadian English (Boberg, 2008), although it can also be found (particularly for /aj/) in some dialects in the United States, especially (but not only) around the Great Lakes region (Vance, 1987; Allen, 1989; Dailey-O'Cain, 1997; Niedzielski, 1999; Labov et al., 2006; Roberts, 2007; Sadlier-Brown, 2012; Stricker et al., 2016; Swan, 2016; Berkson and Herd, 2017; Hamre, 2019). Raising has

[6] This is not to imply that traditional phonological analysis has been unaware of phenomena such as neutralization or varying degrees of predictability of distribution that (as described here) result in partial contrast. But these behavioural results do provide evidence that these phenomena should be understood under the concept of contrast.

even been found in some dialects in England (Britain, 1997; Cardoso, 2015). Historically, the phenomenon has been documented in Canada (specifically Ontario) as far back as people born in 1860 (Chambers, 2006a). Figure 1.6 shows the North American isogloss for Canadian Raising of /aj/ from Labov et al. (2006), extending downward from Canada to include a large portion of the Northern United States. Although their isogloss does not actually include all of Canada,[7] more recent studies that have focused more on Canada have found more consistant raising. Boberg (2008) finds that 88 percent of a regionally-varied sample exhibits a difference of at least 50 Hz in mean F1 values for /aw/, with a comparable 84 percent for /aj/, concluding that Canadian Raising is a "largely uniform feature of Canadian English". Similarly, Hall (2015) finds clear Canadian Raising in both Toronto and Vancouver.

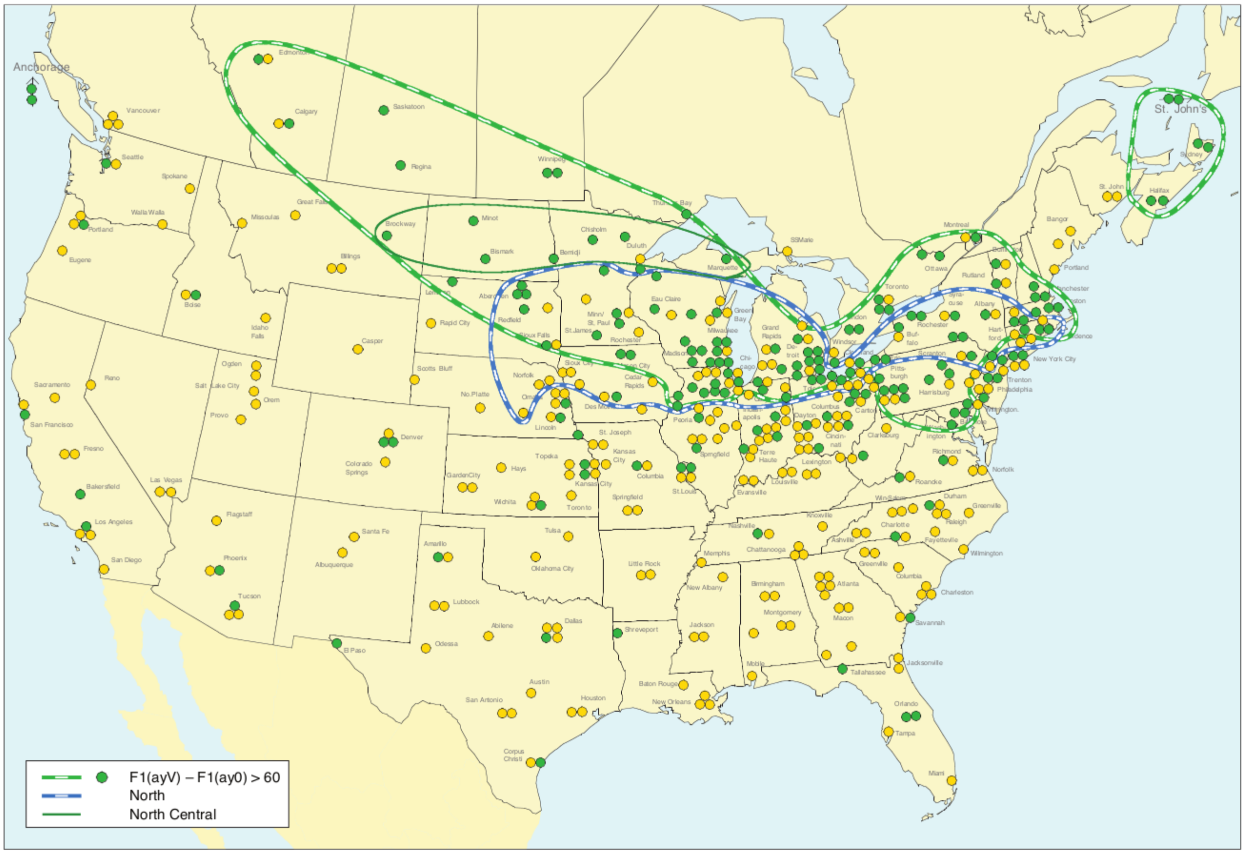

Figure 1.6: Canadian Raising of /aj/ Isogloss from Labov et al. (2006) (their Map 14.10, p. 206)

Despite occurring in some dialects of American English, Canadian Raising plays a prominent role as a stereotype and identifier of Canadian English, especially for /aw/, the diphthong less often raised in American English, whose raised variant is often perceived or exaggerated as *oot and aboot* (Chambers,

[7] Labov et al. (2006) find that Canadian Raising is "not uniform enough to serve as a defining feature of the dialect of Canada", based on the lack of raising in Vancouver, British Columbia in the west and Saint John, New Brunswick in the east.

1973; Boberg, 2008). While the experiences of Joos (1942) almost 80 years ago suggest that non-raising by Americans might have been similarly stereotyped and remarked on by Canadians at the time ("if I use a low diphthong before a fortis consonant [...] the Canadian listener immediately accuses me of drawling", p. 142), more recently it does not appear to be the case that non-raising (for either /aj/ or /aw/) features as prominently in stereotypes of American English by Canadians as raising features in stereotypes of Canadian English by Americans.

1.4.2 Partial Contrast

Although Canadian Raising is canonically an allophonic alternation, with raised diphthongs occurring before voiceless consonants like /t, θ, s/ and non-raised diphthongs occurring before voiced consonants like /d, ð, z/ and elsewhere, its interaction with North American intervocalic /t/–/d/ neutralization results in the possibility of raised/non-raised minimal pairs in one particular phonological environment. The North American flapping rule neutralizes /t/–/d/ to the alveolar flap [ɾ] intervocalically before unstressed vowels (although some sonorants can also precede or follow; see Vaux, 2000) but the raised status of the preceding diphthong determined by the voicing of the consonant remains. This results in [ʌj]~[aj] minimal pairs like [ˈɹʌjɾɪŋ] *writing* (the flap is underlyingly voiceless) versus [ˈɹajɾɪŋ] *riding* (underlyingly voiced flap), [ˈsʌjɾɪŋ] *sighting* versus [ˈsajɾɪŋ] *siding*, and [ˈtʌjɾl̩] *title* versus [ˈtajɾl̩] *tidal.* In principle the same is also possible with [ʌw]~[aw], although actual examples of minimal pairs, such as [ˈklʌwɾɪŋ] *clouting* and *clouding* [ˈklawɾɪŋ], rely on more obscure words that are less recognizable to speakers (the first word in that pair, *clouting*, has various archaic and dialectal meanings). The existence of these minimal pairs in the flap environment, particularly the more common and recognizable minimal pairs like *writing/riding* for the /aj/ diphthong, clearly satisfies the lexical distinction test for contrast as outlined in Hall (2013), although whether it satisfies the predictability of distribution test depends on the level of representation under consideration. The diphthong is predictable from the underlying representation but not from the surface form, making it at least a surface contrast according to this second test. Thus, Canadian Raising is an example of partial contrast. Raised and non-raised diphthongs more resemble contrast (being unpredictable, at least on the surface, and having minimal pairs) in the flap environment, but they more resemble allophony (being predictable and having no minimal pairs) in other environments (Mielke et al., 2008; Hall, 2013; Stevenson and Zamuner, 2017).

Hall (2012) explores the partially contrastive status of [ʌj]~[aj] more deeply with her Probabilistic Phonological Relationship Model (PPRM), which quantifies predictability of distribution on a gradient scale from completely predictable (allophonic) to completely unpredictable (contrastive) and considers environment-specific contrast as well as systemic contrast. As applied to [ʌj] and [aj] (using a corpus of spoken and written Canadian English), she finds that these sounds are close to completely unpredictable in the flap environment, supporting the notion that this environment "considerably disrupt[s] the predictability of these two vowels". However, the flap environment is less common than the other environments, which are predictable, and so overall across environments these two vowel sounds are much closer to the predictable/allophonic side of the gradient scale. This is especially the case when using token-frequency in the corpus instead of type-frequency, suggesting that there are various words that lead to a contrast between [ʌj] and [aj] but they are of relatively low frequency.

One relevant phenomenon separate from, but likely related to, Canadian Raising is [ʌj]-lexicalization. There is evidence from certain dialects, especially in the Great Lakes /aj/ raising region of the United States, that raised [ʌj] is becoming lexicalized and possibly emerging as a new phoneme, which is to say that [ʌj] occurs consistently in certain words that cannot be explained by the Canadian Raising allophonic rule.[8] For example, there is evidence of speakers that have [ʌj] in *spider* and *cider* and an [ʌj]~[aj] minimal pair in *idle/idol* (Fruehwald, 2008; Vance, 1987). *Spider*, *cider*, and *idle* have flaps but (at least based on orthography) they are underlyingly /d/ rather than /t/, and thus we do not expect raised diphthongs here based on Canadian Raising. Interestingly, there is as of yet no evidence of [ʌw] lexicalizing in a similar way. Complicating these findings on [ʌj]-lexicalization somewhat, Hall (2005) recorded speakers from Ontario, finding both unexpected cases of [ʌj] (in non-raising environments) and [aj] (in raising environments) (also found in Ontario and British Columbia by Fullerton, 2019). In most cases these unexpected pronunciations could be explained by the lexical neighbourhood effect and the preceding consonant; salient words with or without raising affect the raising status of other words that share the same onset (e.g., the salient word *meningitis* with allophonic raising causes unexpected raising after /dʒ/ in words like *gigantic* and *angina*). It is unclear whether the lexical neighbourhood effect is an alternative explanation to [ʌj]-lexicalization for these unexpected instances of [ʌj], as opposed to a related phenomenon that influences [ʌj]-lexicalization (helping to explain in which words [ʌj] becomes lexicalized). It is also unclear the extent to which these unexpected pronunciations among speakers in Ontario are the same phenomenon as the unexpected raised variants in the U.S. Great Lakes region, given evidence that the phenomenon varies even within the United

[8] The raised diphthong variants emerging as their own phonemes was actually predicted by Joos (1942), one of the earliest research papers on Canadian Raising.

States (Fruehwald, 2008).

1.5 Online Data Collection

This dissertation makes extensive use of remote web-based data collection. Experiment 1 tests both in-lab and online samples, but all subsequent experiments are online, designed and implemented using the jsPsych JavaScript library (de Leeuw, 2015) and with participants recruited using the Prolific online subject pool (Palan and Schitter, 2018). Online data collection has been found to be an effective alternative to the in-lab testing of primarily undergraduate participants that has traditionally been predominant in psycholinguistics and psychology; benefits include the potential for larger and more diverse sample sizes and faster data collection (Buhrmester et al., 2011; Mason and Suri, 2012; Paolacci and Chandler, 2014; Buhrmester et al., 2018). Gosling et al. (2004) found that web-based experimental results are consistent with findings from traditional methods of data collection and are not negatively affected by nonserious responders, Sprouse (2011) found that results from Amazon's Mechanical Turk are "almost indistinguishable from laboratory data", and Hauser and Schwarz (2016) found that participants recruited from Amazon's Mechanical Turk are actually more attentive than traditional subject pools.

1.6 Dissertation Outline

The primary focus of this dissertation is the kind of context-based partial contrast effect found in Murphy et al. (2016) and Celata (2008), where participants respond differently to partially contrastive sound pairs in contrastive environments than in non-contrastive (allophonic or neutralizing) environments. In the case of Canadian Raising, this means responding differently to raised and non-raised diphthong pairs ([ʌj]~[aj] and [ʌw]~[aw]) in the contrastive (flap) environment than in the other allophonic environments. Based on findings from research on categorical perception that listeners more accurately and more quickly discriminate between sounds that are contrastive in their language than sounds that are non-contrastive (e.g., Liberman et al., 1957; Lago et al., 2015; Boomershine et al., 2008), and findings from the research on partial contrast, in particular the finding from Western Tuscan from Celata (2008) of faster and more accurate identification of partially contrastive consonants in the contrastive environment than in the neutralizing environment, the primary hypothesis for Canadian Raising is that discrimination of raised and non-raised vowels will be more accurate in the contrastive (flap) environment than in the other allophonic environments. Discrimination will be tested using an AXB discrimination paradigm, a variation of the ABX discrimination task (Munson and Gardner, 1950) where the sound to be judged is presented in between, rather than after, the sounds that participants must compare to.

This introduction is Chapter 1 of this dissertation. Chapter 2 takes a more detailed look at the phonetic properties of Canadian Raising, providing a review of past phonetic analyses of the diphthong variants as well as a new phonetic analysis of the recordings done for the stimuli in this dissertation, in order to better understand the stimuli for the results of the following experiments.

The above-mentioned context-based partial contrast discrimination effect for Canadian Raising is investigated in Chapters 3, 4, and 5. Chapter 3 presents the initial findings on partial contrast and Canadian Raising. Within this chapter, Experiment 1 finds a context-dependent discrimination effect for the /aj/ diphthong but not the /aw/ diphthong using real word stimuli, while Experiment 2 replicates this design using non-word stimuli to clarify the reason for the diphthong discrepancy, finding that it was not only a result of the clearer minimal pairs (like *writer/rider*) that are available for /aj/. Chapter 4 contributes an additional set of discrimination experiments that seek to understand the role of the lexical items (minimal pairs) that are available in explaining the unexpected finding in the initial experiment. Within this chapter, Experiment 3 tests the effect of lexicality using a more similar design to the previous two experiments, while Experiment 4 tests this effect using a more novel design. Next, Chapter 5 provides a series of semi-replications of these past experiments (Experiments 1b, 1c, 2b, 3b, and 4b), which demonstrate that the context-dependent discrimination effect that has been investigated and understood in the previous experiments actually depends to a large extent on the listeners having a certain degree (or type) of experience with the stimuli and/or speakers.

Chapter 6 moves beyond this context-based partial contrast effect investigated in previous chapters, presenting a final experiment (Experiment 5) that compares speakers of Canadian English to speakers of American English from two regions: the U.S. North (Great Lakes) raising region, as well as the U.S. West region where raising is not traditionally expected. The findings from a short dialect survey included in this experiment provide some concrete numbers on the prevalence of Canadian Raising and [ʌj]-lexicalization in different regions of North America, while the discrimination results have relevance for (among other topics) cross-dialectal perception and the interaction between stereotypes, production, and exposure as they affect perception.

Finally, Chapter 7 provides a summary and general discussion of the findings from Chapters 2 to 6, cover-ing four main areas of findings: (partial) contrast, the perceptual basis for the emergence of /ʌj/ as a new phoneme, the effect of dialectal stereotypes and exposure on perception, and online data collection.

Chapter 2

Canadian Raising Diphthongs

This dissertation is composed of five perception experiments (with an additional five semi-replications of those experiments) that use an AXB paradigm to test discrimination of raised and non-raised diphthongs. All experiments test [ʌj]~[aj], and Experiments 1, 2, and 5 additionally test [ʌw]~[aw]. Two native speakers of Canadian English in their mid-to-late 20s (one female from Ontario and one male from Nova Scotia) were recorded for all of the stimuli in this dissertation. This chapter presents a phonetic analysis of the Canadian Raising diphthongs used as stimuli in this dissertation, to provide context for the perception experiments and results that will follow in later chapters. Most notably, this phonetic analysis finds a greater difference between raised and non-raised variants of /aj/ than for /aw/, which will be relevant for discrimination differences between /aj/ and /aw/ that are found throughout the experiments.

2.1 Previous Studies

As seen in Chapter 1, Canadian Raising is typically defined in terms of the diphthong onset (higher in voiceless contexts than voiced contexts), as schematized in Figure 2.1. For example, Labov et al. (2006) set a benchmark for Canadian Raising as a 60 Hz or greater difference in F1 values (corresponding to vowel height) in the diphthong onset between voiceless and voiced environments. However, other differences between raised and non-raised diphthongs have been documented. This includes fronting for the raised variant of /aw/ (Chambers and Hardwick, 1986; Boberg, 2008; Hall, 2015), and also a difference in the diphthong offset, particularly for /aj/ (Rosenfelder, 2007; Hall, 2015), as seen in Figure 2.2 from Hall (2015).

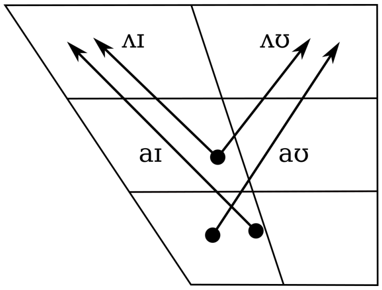

Figure 2.1: Traditional depiction of Canadian Raising (from Wikimedia Commons, based on Henry Rogers, The Sounds of Language: An Introduction to Phonetics, 2000, p. 124)

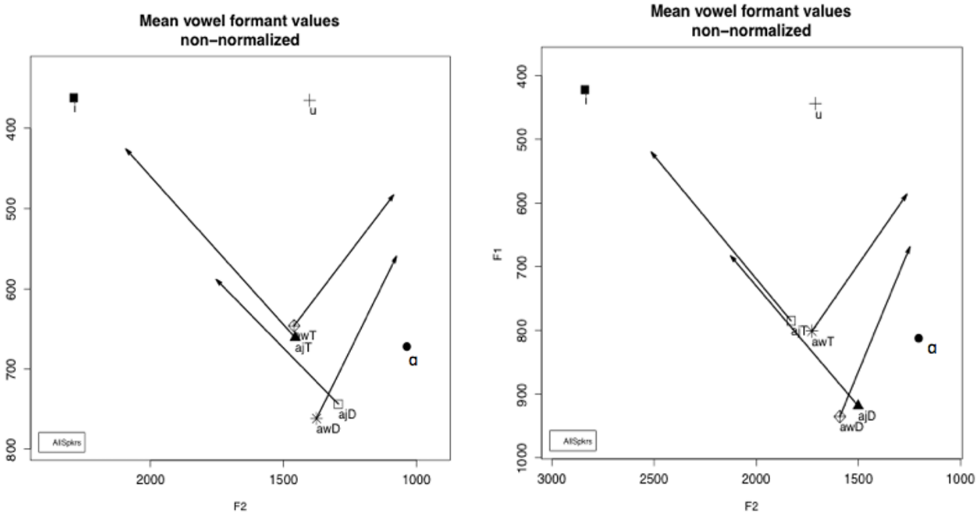

Figure 2.2: Raised and non-raised diphthongs (L: males, R: females) from Hall (2015) (her Figure 5)

2.2 Phonetic Analysis of Stimuli

2.2.1 Rationale

This phonetic analysis covers the stimuli used in Experiments 1, 2, and 5, which are the experiments that include both the /aj/ and /aw/ diphthongs. Beyond providing a general sense of the vowels that the participants encountered, this phonetic analysis has two specific goals: to compare the stimuli against the definition of Canadian Raising from Labov et al. (2006),[1] and to determine whether there is a greater difference between raised and non-raised diphthongs for /aj/ than /aw/ (to understand the discrimination advantage for /aj/ over /aw/ consistently found in these experiments).

[1] This is calculated for informational purposes to describe the stimuli using the criterion for Canadian Raising from the *Atlas of North American English* (Labov et al., 2006). However, it was not a component of initial stimulus selection. As mentioned (see Figure 2.2), a vowel height difference at onset is only one of the differences between raised and non-raised diphthongs.

2.2.2 Stimulus Creation

The process of stimulus creation was comparable across all experiments (with differences in the contexts and diphthongs that were included in each experiment). For each experiment, word pairs like *sight* and *side* were recorded by both the male and female speaker. *Sight* has /aj/ in a voiceless context, which results in a raised diphthong [ʌj], while *side* has /aj/ in a voiced context, resulting in a non-raised [aj] realization. These vowels were extracted and spliced into a different recording of *sight* to make two pronunciations of that word: one with a raised vowel [ˈsʌjt] and one with a non-raised vowel [ˈsajt] (both vowels were modified in duration to match the duration of the original raised vowel that they replaced). Depending on the experiment, the same vowel tokens were also used to create two pronunciations of *side* and two pronunciations of *sighting*, so that discrimination of raised and non-raised vowels could be tested in the voiceless, voiced, and flap contexts. Figure 2.3 provides a schematization of the stimulus creation process. It is the raised and non-raised vowel tokens taken from words like *sight* and *side* and spliced into other recordings that will be analyzed in this chapter.

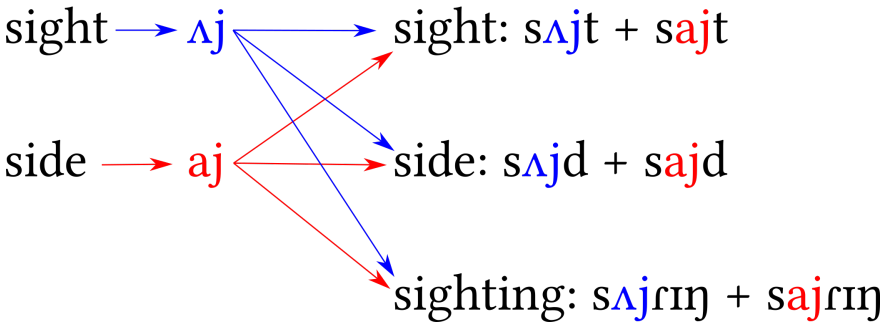

Figure 2.3: Schematization of cross-splicing of vowel tokens

2.2.3 Method

2.2.3.1 Items

For the Experiment 1 recordings, analysis was performed on the real words that were recorded and annotated in the process of stimulus creation for Experiment 1: *sight* and *side* (for /aj/), as well as *doubt*, *endowed*, *clout*, and *cloud* (for /aw/). These are from the voiceless and voiced conditions only (excluding the flap condition) because all of the raised and non-raised vowels that were used came from those conditions. Two other words were recorded for /aj/ (*write* and *ride*) but they were not included in this analysis because the liquid onset and the diphthong were kept together and treated as one unit for manipulation in stimulus creation. Only the vowels that were used as stimuli were included in this analysis (for each speaker, two recordings of each of the above mentioned six words). Waveforms and spectrograms for a raised and non-raised instance of /aj/ are presented in Figure 2.4 for the female speaker and Figure 2.5 for the male speaker.

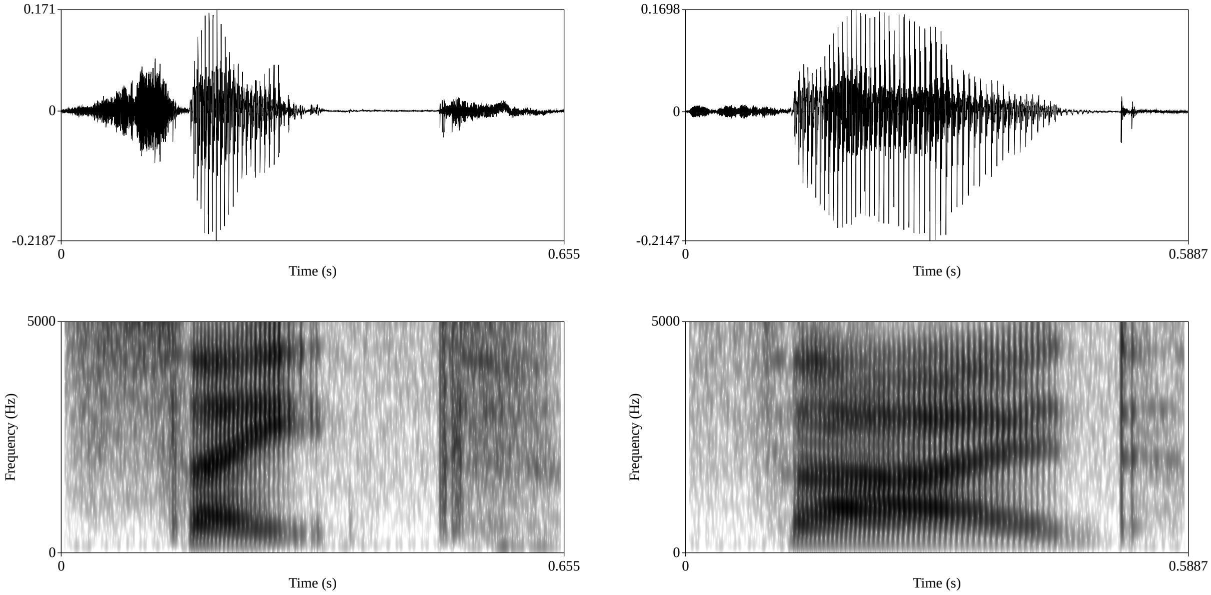

Figure 2.4: Waveform and spectrogram for 'sight' (L) and 'side' (R) from female speaker in Experiment 1 recordings

For the Experiment 2 recordings, analysis was performed on the non-words that were recorded and annotated in the process of stimulus creation for Experiment 2: *fidight*, *fidide*, *kuvight*, *kuvide*, *stazight*, and *stazide* (for /aj/) and *fidaut*, *fidaud*, *kuvaut*, *kuvaud*, *stazaut*, *stazaud* (for /aw/). For each speaker there were two recordings of each of those 12 words used as stimuli and thus analyzed here.

For the Experiment 5 recordings, analysis was performed on the real words that were recorded and annotated in the process of stimulus creation for Experiment 5: *height* and *hide* (for /aj/), as well as *out* and *how'd* (for /aw/). For each speaker there were five recordings of each of those four words.

2.2.3.2 Procedure

These recordings were previously annotated for stimulus creation in Experiments 1, 2, and 5 using TextGrid in Praat, separating the diphthong from the initial and following sounds (e.g., "sight" [ˈsʌjt] was divided

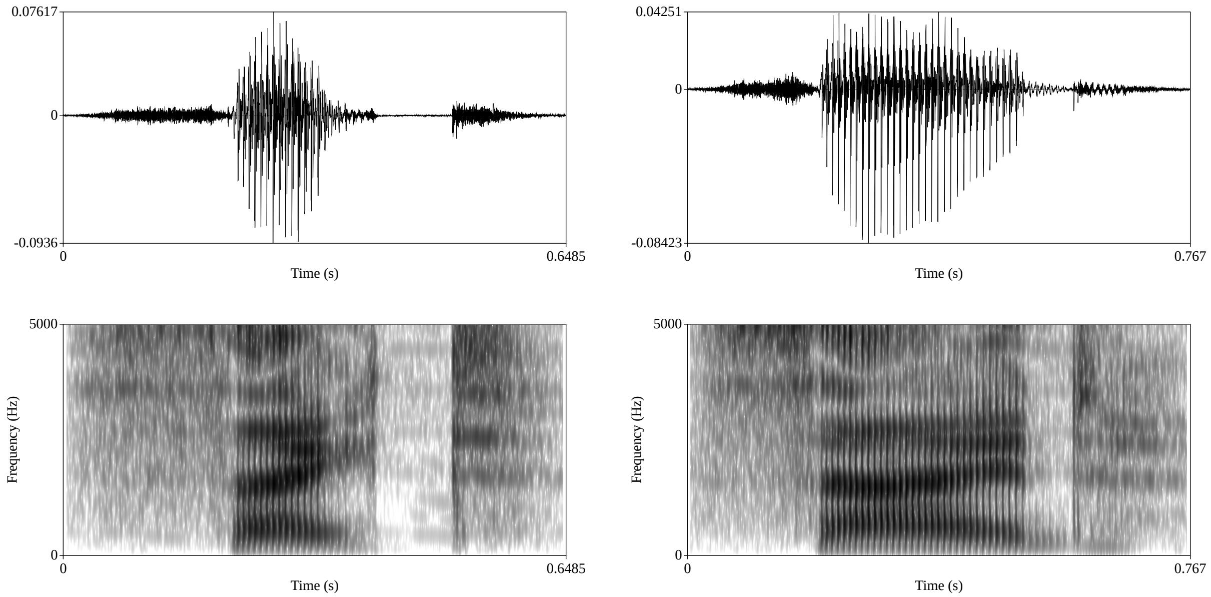

Figure 2.5: Waveform and spectrogram for ‘sight’ (L) and ‘side’ (R) from male speaker in Experiment 1 recordings

into [s] + [ʌj] + [t]). Due to the generally clear formant patterns in these recordings, as seen in Figure 2.4 and Figure 2.5, a Praat script was used to automatically identify and record the F1 and F2 values for six points (0 percent, 20 percent, 40 percent, 60 percent, 80 percent, and 100 percent) of the diphthong. These results were manually checked for outliers or unexpected values and in a small number of cases replaced with a value recorded manually.

The first analysis of the resulting formant values will determine whether these speakers exhibit a difference of at least 60 Hz in mean F1 values (i.e., vowel height) in the diphthong onset between voiceless and voiced environments to meet the criterion for Canadian Raising from Labov et al. (2006). This will be tested by comparing the mean F1 values of raised and non-raised vowels for each diphthong at the 20 percent point. The second analysis of these vowels will be to determine whether there is a greater difference between raised and non-raised variants of /aj/ than /aw/. This will involve both F1 and F2 (converted from Hz to the Bark scale), and it will be calculated as the average Euclidean distance between raised and non-raised vowels across the four middle points (20 percent, 40 percent, 60 percent, and 80 percent). This involved calculating the Euclidean distances at each of the four middle points and then averaging them (separately for each speaker and diphthong). The choice to include these four middle points was made as a result of general findings that sampling multiple points from a diphthong is preferred over two (Hall, 2015; Fox and Jacewicz, 2009; Davidson, 2006), and specific findings in Hall (2015) that raised and non-raised diphthongs

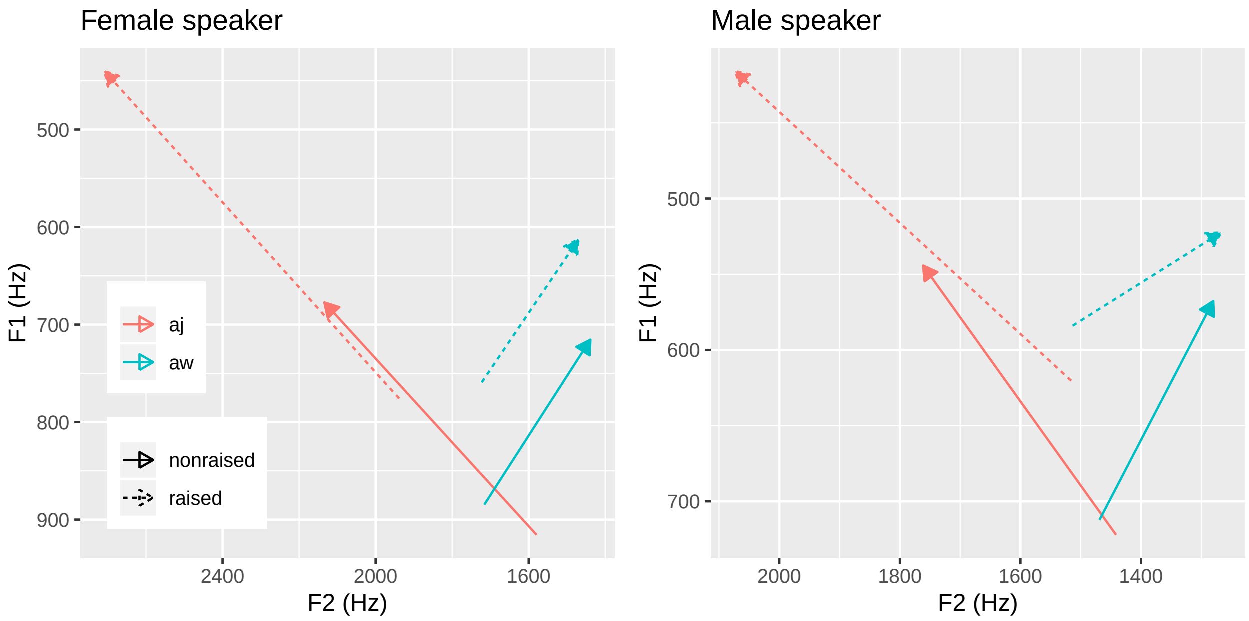

Figure 2.6: Mean formant values at 20% and 80% from Experiment 1 recordings

in Canadian English vary not just in their onset but also their offset or glide segment.

2.2.4 Results

2.2.4.1 Experiment 1 Recordings

The mean F1 and F2 values for the 20 percent and 80 percent points for both speakers are provided in Figure 2.6, separated by diphthong and raising (an additional visualization of these vowels, in a format more resembling a spectrogram, is provided in the appendix in Figures A.1 and A.2). Table 2.1 shows the difference in F1 between raised and non-raised vowels at 20 percent. For both diphthongs, both speakers exceed the 60 Hz Labov et al. (2006) cutoff for Canadian Raising by a substantial margin (an average of 124 Hz). To quantify the overall difference between raised and non-raised variants for /aj/ and /aw/, the average Euclidean distance (using the Bark scale) between raised and non-raised (across the four middle points: 20 percent, 40 percent, 60 percent, and 80 percent) was 2.69 for the female speaker's /aj/ compared to 0.99 for her /aw/, and 1.31 for the male speaker's /aj/ compared to 0.69 for his /aw/. Thus, the difference between raised and non-raised variants was larger for /aj/ than for /aw/ for both speakers in these recordings (and the female speaker had a greater difference between raised and non-raised vowels than the male speaker did).

Table 2.1: F1 values and differences between onset (20 percent) of raised and non-raised vowels in Experiment 1 recordings

speaker	diphthong	nonraised	nr.sd	raised	r.sd	difference
female	aj	916	23	776	2	-140
female	aw	885	33	759	20	-125
male	aj	722	10	620	14	-102
male	aw	712	27	584	55	-128

Table 2.2: F1 values and differences between onset (20 percent) of raised and non-raised vowels in Experiment 2 recordings

speaker	diphthong	nonraised	nr.sd	raised	r.sd	difference
female	aj	851	36	762	58	-89
female	aw	754	214	733	9	-22
male	aj	660	24	575	29	-86
male	aw	656	62	557	25	-98

2.2.4.2 Experiment 2 Recordings

The mean F1 and F2 values for the 20 percent and 80 percent points for both speakers are provided in Figure 2.7, separated by diphthong and raising (an additional visualization of these vowels, in a format more resembling a spectrogram, is provided in the appendix in Figures B.1 and B.2). Table 2.2 shows the difference in F1 between raised and non-raised vowels at 20 percent. Overall the speakers exceeded the 60 Hz Labov et al. (2006) cutoff for Canadian Raising with an average of 74 Hz, but the female speaker's /aw/ tokens exhibited a smaller difference in mean F1 at 20% (22 Hz). To quantify the difference between raised and non-raised variants for /aj/ and /aw/, the average Euclidean distance (using the Bark scale) between raised and non-raised vowels was 1.62 for the female speaker's /aj/ compared to 1.24 for her /aw/, and 1.46 for the male speaker's /aj/ compared to 0.80 for his /aw/. Thus, the difference between raised and non-raised variants was larger for /aj/ than for /aw/ for both speakers in these recordings (and the female speaker had a greater difference between raised and non-raised vowels than the male speaker did).

2.2.4.3 Experiment 5 Recordings

The mean F1 and F2 values for the 20 percent and 80 percent points are provided for both speakers in Figure 2.8, separated by diphthong and raising (an additional visualization of these vowels, in a format more resembling a spectrogram, is provided in the appendix in Figures C.1 and C.2). Table 2.3 shows the

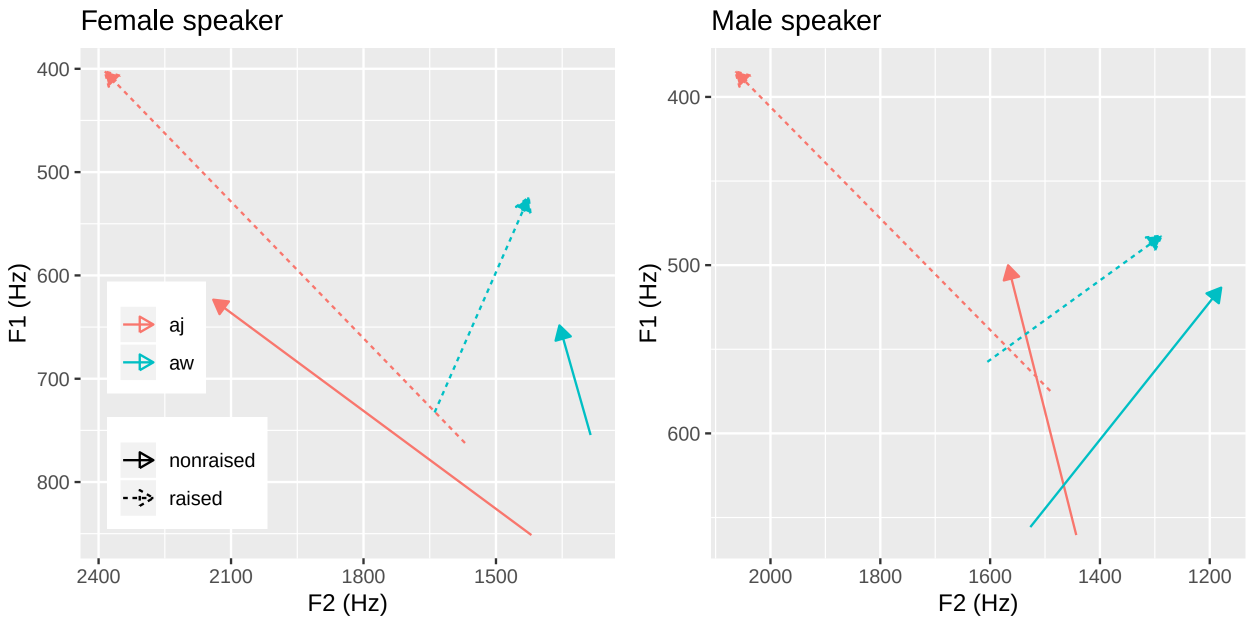

Figure 2.7: Mean formant values at 20% and 80% from Experiment 2 recordings

Table 2.3: F1 values and differences between onset (20 percent) of raised and non-raised vowels in Experiment 5 recordings

speaker	diphthong	nonraised	nr.sd	raised	r.sd	difference
female	aj	943	12	878	36	-66
female	aw	999	27	849	24	-151
male	aj	743	27	724	35	-19
male	aw	760	33	614	16	-147

difference in F1 between raised and non-raised vowels at 20 percent. Overall the speakers exceeded the 60 Hz Labov et al. (2006) cutoff for Canadian Raising with an average of 96 Hz, but the male speaker's /aj/ tokens exhibited a smaller difference in mean F1 at 20% (19 Hz). To quantify the difference between raised and non-raised variants for /aj/ and /aw/, the average Euclidean distance (using the Bark scale) between raised and non-raised vowels was 2.33 for the female speaker's /aj/ compared to 0.94 for her /aw/, and 1.66 for the male speaker's /aj/ compared to 1.19 for his /aw/. Thus, the difference between raised and non-raised variants was larger for /aj/ than for /aw/ for both speakers in these recordings (and the female speaker had a greater difference between raised and non-raised vowels than the male speaker for /aj/, but the male speaker had a greater difference for /aw/).

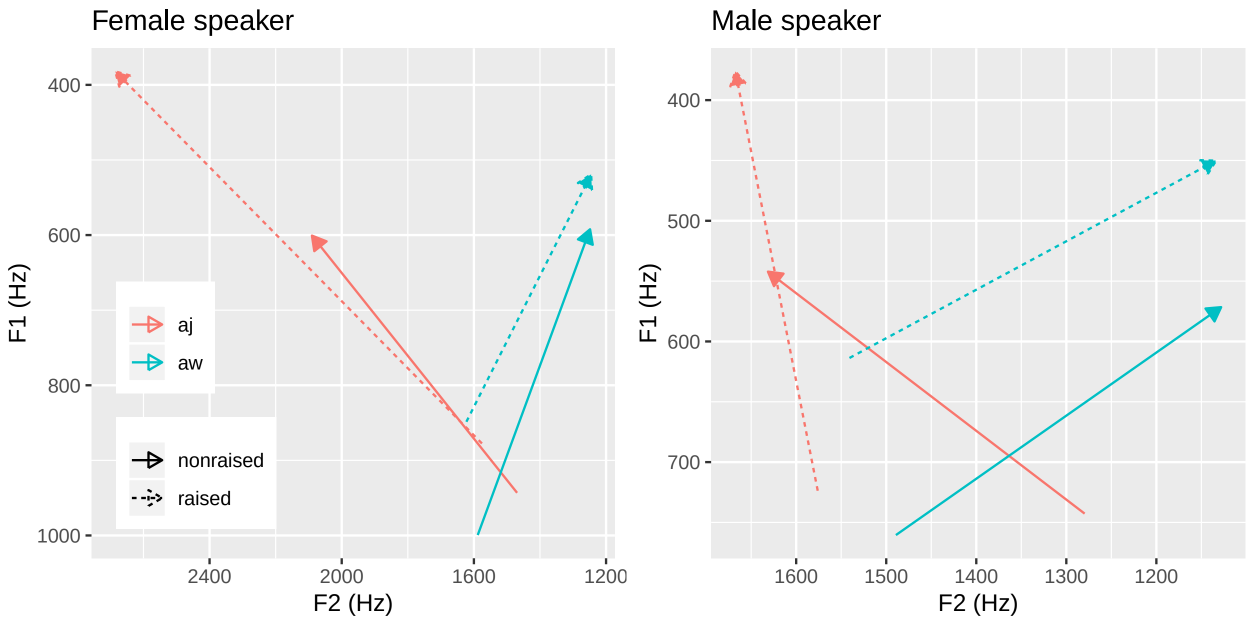

Figure 2.8: Mean formant values at 20% and 80% from Experiment 5 recordings

2.2.5 Discussion

Overall these tokens exhibited Canadian Raising as defined by Labov et al. (2006) across the recordings for all three experiments, although there were isolated token averages that did not meet this threshold (the female speaker's /aw/ in Experiment 2 and the male speaker's /aj/ in Experiment 5). The token averages that did not meet this threshold should not be interpreted as having no difference between raised and non-raised diphthongs, because this does not take into account F1 differences at the diphthong offset or F2 differences at either the onset or offset.

Regarding the comparison between diphthongs, there was a greater phonetic difference between raised and non-raised variants of /aj/ than /aw/. Using the Bark scale, the average Euclidean distance between raised and non-raised diphthongs (across both speakers) was 2.00 for /aj/ compared to 0.84 for /aw/ in the Experiment 1 recordings, 1.54 for /aj/ compared to 1.02 for /aw/ in the Experiment 2 recordings, and finally 2.00 for /aj/ compared to 1.06 for /aw/ in the Experiment 5 recordings. While this covers only two speakers, and in fact only the subset of their vowel recordings (only the ones that were used for the stimuli in Experiments 1, 2, and 5), similar results can be seen in Hall (2015), a much more comprehensive phonetic analysis that included sixty speakers, evenly distributed between Toronto and Vancouver and between male and female. A manual calculation of the values in her Table 3 provides an average Euclidean distance between raised and non-raised diphthongs (again using the Bark scale) as 1.53 for /aj/ and 0.83 for /aw/.[2] This finding of a greater phonetic difference for [ʌj]~[aj] than [ʌw]~[aw] will be relevant for the perception results of Experiments 1, 2, and 5.

The diphthong productions of the two speakers recorded for the stimuli in this dissertation resemble the productions recorded and analyzed by Hall (2015) in other ways as well, such as a higher ending position of the raised variants compared to their non-raised counterparts, especially for /aj/.

[2] This calculation was based on two points from the diphthong (onset and glide) rather than four, and conversion from Hertz to Bark took place on the group averages (e.g., average F1 of Toronto males for the onset of [ʌj]) rather than the individual tokens (the conversion from Hertz to Bark is nonlinear, and so the results are somewhat different depending on whether the mean is calculated before or after conversion). Still, these figures show a greater difference between raised and non-raised variants of /aj/ than /aw/ in her larger dataset.

Chapter 3

Basic Findings

This chapter presents the first two experiments of this dissertation, which provide the basic findings about the context-based effect of partial contrast on discrimination that the following two chapters build on and clarify. Experiment 1 investigates whether discriminability of raised and non-raised diphthongs in Canadian English varies according to the contrastive status of the phonological environment (following consonant), finding that it does—but only for one of the two diphthongs involved in Canadian Raising (/aj/ but not /aw/). Experiment 2 extends the design of Experiment 1 to non-words, finding a comparable (although weaker) effect of partial contrast on discrimination for /aj/ (but not /aw/), suggesting the diphthong difference found in Experiment 1 was in fact related to the diphthongs themselves and not just the lexical items used.

3.1 Experiment 1

3.1.1 Rationale

This experiment tests the ability of speakers of Canadian English to discriminate raised and non-raised diphthongs ([ʌj]~[aj] and [ʌw]~[aw]) in three distinct phonological environments:

1. Before a voiceless sound (/t/), which licenses raising.
2. Before a voiced sound (/d/), which does not license raising.

3. Before a flap ([ɾ]), which variably licenses raising (depending on whether the flap is underlyingly a /t/ or a /d/), which can lead to minimal pairs like *writing/riding*.

The hypothesis, in line with findings in Whalen et al. (1997), Boomershine et al. (2008), Celata (2008), and Barrios et al. (2016), is that discrimination of the diphthong variants will be better in the environment where they can create different words (before a flap) than the environments where they cannot: before a (non-flap) voiceless or voiced sound.

This experiment was tested both in-lab, on a sample of primarily undergraduate students at the University of Toronto, and online, using the Prolific online subject pool (Palan and Schitter, 2018) to access a broader sample of speakers of Canadian English. Experiment medium (in-lab and online) will be included in the analysis; if there are no critical differences, this would strengthen the generalizability of the findings and support the legitimacy of browser-based online experiments, which are used for the rest of the dissertation.

3.1.2 Method

3.1.2.1 Participants

Thirty-two native speakers of Canadian English took part in this experiment at the University of Toronto Phonetics Lab. For participation in this and a different experiment in the same session, they were reimbursed with either CAD$10 or course credit. In addition, forty-seven native speakers of English (born and currently living in Canada, mean age = 30, SD = 8.9, 25 men and 22 women) were recruited to participate in a browser-based online experiment through Prolific (an additional three participants were tested but excluded from analysis for reasons of language background or country status). The geographic distribution of online participants (using province of birth) is provided in Figure 3.1. Based on the distribution of English speakers (by mother tongue) in Canada in Figure 3.2, for this experiment and others going forward we generally expect a majority of participants to come from Ontario (which has 46 percent of Canada's English speakers), British Columbia (16 percent), and Alberta (15 percent). Quebec is the second most populated province, but it is primarily francophone and thus it only has 3 percent of Canada's English speakers by mother tongue. Provincial origins and demographic characteristics (age, sex) were not recorded for the in-lab participants. Most online participants indicated little-to-no experience with linguistics. Participants provided informed consent using the same consent form (online or offline). Online participants were paid CAD$5.20 for a session of approximately 20 minutes. In total, there were 78

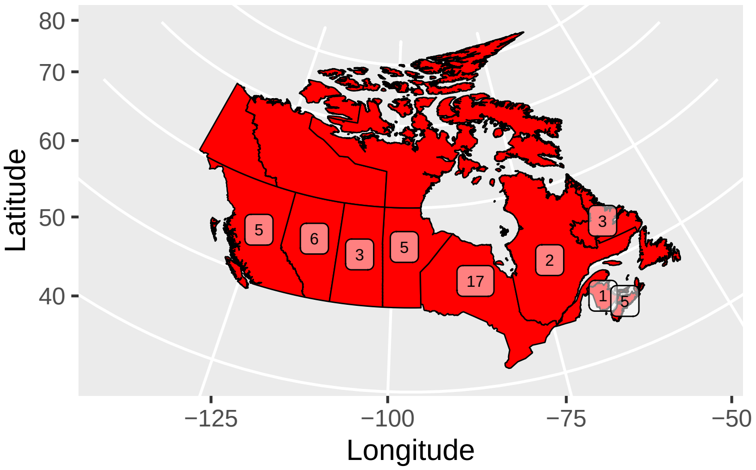

Figure 3.1: Distribution of Online Participants in Experiment 1 (province of birth)

Table 3.1: Experiment 1 stimuli

Voiceless	Flap (Raised)	Flap (Non-raised)	Voiced
write	writing	riding	ride
sight	sighting	siding	side
clout	clouting	clouding	cloud
doubt	doubting	Dowding	endowed

participants in this experiment.

3.1.2.2 Items

The stimuli in this experiment are the 16 words with an /aj/ or /aw/ diphthong in Table 3.1. There were four words used to test discrimination of raised and non-raised diphthongs in the voiceless environment; each word was created with raised and non-raised versions, e.g., *write* as [ˈɹʌjt] and [ˈɹajt]. There were also four words used to test discrimination of diphthong variants in the voiced environment, with *ride* being created as [ˈɹʌjd] and [ˈɹajd]. There were eight words in the flap environment (separated into two columns) not because more stimuli were used, but because the words were chosen with the intention that the raised and non-raised versions would be interpreted as two different lexical items. Therefore, while [ˈɹʌjt] and [ˈɹajt] are both intended to be interpreted as *write*, and [ˈɹʌjd] and [ˈɹajd] are both intended to be interpreted as *ride*, their equivalents in the flap environment—[ˈɹʌjɾɪŋ] and [ˈɹajɾɪŋ]—were intended to be interpreted as *writing* and *riding*, respectively. This lexical contrast in the flap environment was intended to apply to both /aj/ words (top two rows in Table 3.1) and /aw/ words (bottom two rows), but due to the

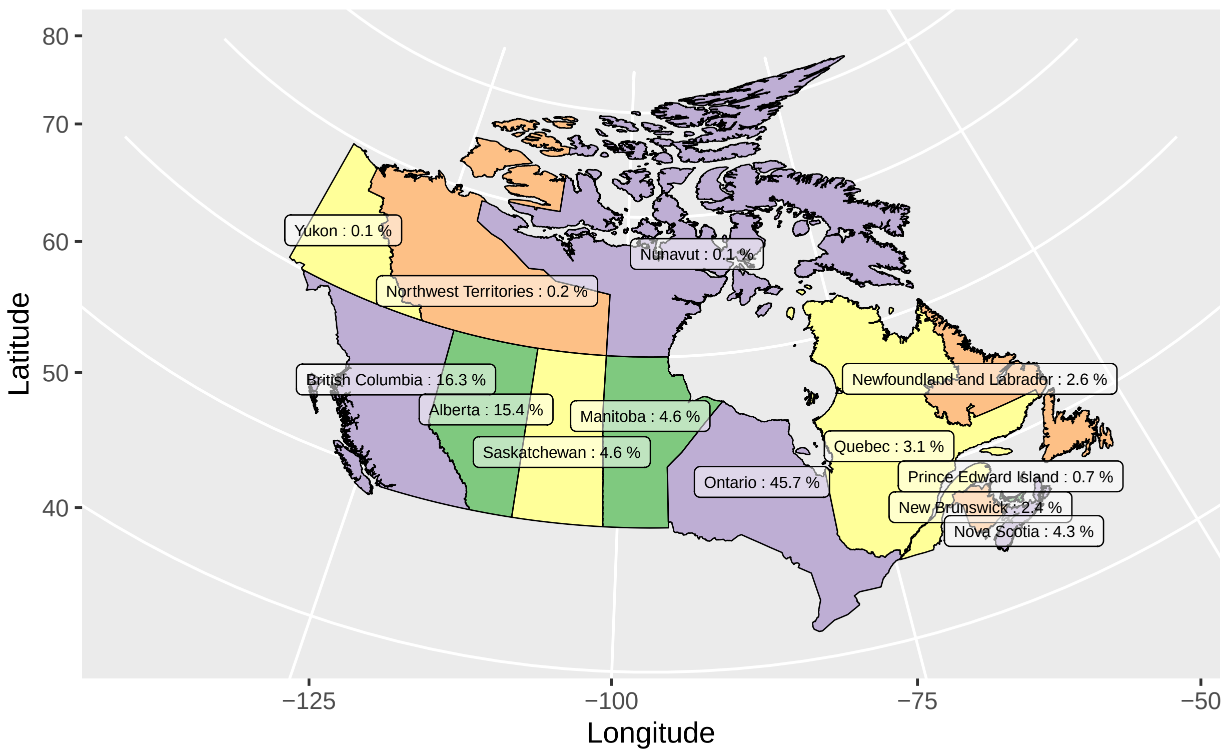

Figure 3.2: Canadian provinces and territories with their share of Canada's English speakers by mother tongue (i.e., the numbers add up to 100 percent), 2016 census

lack of suitable /aw/ words, the lexical distinctions were more dubious: *clouting* (archaic/uncommon) ~ *clouding*, and *doubting* ~ *Dowding* (a surname). This will be relevant in the results.

For the purposes of stimulus creation, the words were organized such that one word in each environment was matched with a word in the other environments according to the onset preceding the vowel (/ɹ, s, kl, d/—see rows in Table 3.1). This means that within each set, the same vowel tokens could be used for the voiceless, voiced, and flap environments (in other words, splicing the same vowel tokens between different onsets was avoided). Two speakers of Canadian English, one male (from Nova Scotia) and one female (from Ontario), both in their mid-to-late 20s, were recorded for creation of the stimuli. Recording took place at the Phonetics Lab at the University of Toronto, recorded at 48 KHz and 24-bit using a Sound-Devices 722 digital audio recorder. They were instructed to read each of the words naturally, except to clearly produce the final /t/ and /d/ sounds in the voiceless and voiced words, to avoid ambiguity (their natural pronunciations of /t/ tended to be unreleased or a glottal stop, while /d/ was often devoiced). The recordings were annotated using TextGrid in Praat (Boersma and Weenick, 2017), dividing each word at zero crossings before and after the diphthongs, e.g., [s] + [ʌj] + [t], except for the series of words beginning with /ɹ/. Due to the lack of clear boundry between the liquid and the following vowel, that series

Table 3.2: Experiment 1 stimulus vowel durations (ms)

Gender	Voiceless	Flap	Voiced
Male	234.8	218.8	361.2
Female	177.3	185.1	352.0
Average	206.0	202.0	356.6

was only divided into two parts (e.g., [ɹʌj] + [t]) and so the /ɹ/ and the diphthong were treated as one unit for cross-splicing and manipulation.

The process for cross-splicing and manipulation happened as follows. For each gender's recording of each series of words (e.g., the series of words with an /s/ onset), two instances of a raised vowel were extracted from separate recordings of the voiceless word (*sight*), and two instances of a non-raised vowel were extracted from separate recordings of the voiced word (*side*). These vowels were then spliced into two (different) recordings of the voiceless word *sight*, two (different) recordings of the voiced word *side*, and two recordings of the flap word *sighting/siding*.[1] This results in four stimuli files for *sight* for each gender (two with a raised vowel and two with a non-raised vowel), and the same for the voiced (*side*) and flap (*sighting/siding*) environments. The process is schematized in Figure 2.3 in Chapter 2.

When the vowels were extracted and spliced into different recordings, their duration was modified to match the original vowel duration of that individual recording. This means that the raised and non-raised vowels spliced into *sight* have the same duration, which matches the duration of the original (raised) vowel in that recording of *sight*. Duration manipulation was done using the Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) method by means of the Praat Vocal Toolkit (Corretge, 2012). As can be seen in Table 3.2, which shows the durations of the vowels (this includes the onset only for the set of words with /ɹ/) in each environment, the vowels in the voiced environment were notably longer than those in the voiceless or flap environments, consistent with past findings on the effect of consonant voicing on preceding vowel duration in English (Chen, 1970; Cho, 2016).[2]

The stimuli were normalized to an average root mean square intensity of 70 dB Sound Pressure Level (SPL).

[1] One of the tokens from the flap environment came from the recording of the underlyingly voiceless word (e.g., *sighting*) and one came from the underlyingly voiced word (e.g., *siding*). This is because Braver (2013) finds that /t, d/ ⟶ [ɾ] is actually incomplete neutralization; vowels preceeding /t/ flaps are slightly shorter in duration (9 ms) than vowels preceeding /d/ flaps, although Braver (2014) finds this difference to be imperceptible.

[2] Broken down by diphthong rather than gender, there was a pattern of /aj/ diphthongs being longer in duration than /aw/: 229 ms versus 183 ms in the voiceless context, 211 ms versus 192 ms in the flap context, and 367 ms versus 346 ms in the voiced context. Across all of these contexts, /aj/ diphthongs were on average 29 ms longer.

The final result was 96 sound files: 12 words × two speakers × two variants of the diphthong (raised and non-raised) × two versions.

3.1.2.3 Procedure

This experiment was an AXB discrimination task. For each trial the participant heard three words (A, X, and B). Of the first and third words (A and B), one had a raised vowel and one had a non-raised vowel. The middle word (X) had a vowel that matched A or B, and participants had to indicate whether the middle word sounded more like the first or more like the last, using "z" and "m" on the keyboard (for the in-lab participants) or using their mouse to click a button on the screen (for the online participants). An example of a trial for *write* is "[ˈɹʌɪt] ... [ˈɹʌɪt] ... [ˈɹaɪt]" (answer: first), for *ride* is "[ˈɹʌɪd] ... [ˈɹaɪd] ... [ˈɹaɪd]" (answer: second), and for *writing/riding* is "[ˈɹaɪɾɪŋ] ... [ˈɹaɪɾɪŋ] ... [ˈɹʌɪɾɪŋ]" (answer: first). Participants always compared sounds across voices; either A and B were male and X was female, or vice versa.

The in-lab version of the experiment was designed and run in the OpenSesame experiment builder software (Mathôt et al., 2012). Participants were tested on a computer with headphones in a quiet room. The inter-stimulus interval (ISI) between each word within a trial was 1000 ms, in order to make use of a more phonetic or phonemic (as opposed to auditory) level of processing (Werker and Logan, 1985). In total there were 192 trials, with 64 from each condition (voiceless, voiced, and flap) interspersed. The instructions explained that one of the words that they would hear is a surname (*Dowding*). This experiment took approximately 20 minutes.

The online version of the experiment implemented this same design using the jsPsych JavaScript library (de Leeuw, 2015). The trial structure remained the same (192 total trials, including 64 voiceless, 64 voiced, and 64 flap trials). Due to the mechanics of jsPsych, the breaks between words in a trial were programmed as a stimulus onset asynchrony (SOA)—the time between the beginning of the previous stimulus and the beginning of the next—instead of an inter-stimulus interval (ISI), which is the time between the *end* of the previous stimulus and the beginning of the next. An SOA of 1500 ms was chosen to approximate the ISI of 1000 ms in the OpenSesame version of the experiment.

3.1.3 Results

3.1.3.1 Mixed Effects Model

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 3.3. The reference level (flap environment) has been placed in the middle for comparison with the other two environments. Visually, there is an advantage for the voiced and flap environments over the voiceless environment for the /aj/ diphthong and an advantage for the voiced environment over the voiceless and flap environments for the /aw/ diphthong. Response time was not included in the analysis (or any analysis until Experiment 5) because the length of the stimuli is not the same across contexts (the voiced context has a longer vowel than the voiceless context, while the flap context is longer than the voiceless context due to the additional syllable).

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The fixed effects were environment (three levels: voiceless, voiced, and flap), diphthong (two levels: aj and aw), and medium (two levels: lab and online). The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level for environment, diphthong, and medium were "flap", "aj", and "lab", respectively. The approach to random effects was to use the maximal structure justified by the experimental design that does not result in a failure to converge or a singular fit (as judged by asSingular() from lme4) (Barr et al., 2013). The result was by-subjects and by-items random intercepts, as well as by-subjects random slopes for environment. The items were the rows in Table 3.1, which shared the same onset and the same vowel tokens. The results of the mixed effects analysis are presented in Table 3.3.

To summarize the significant findings ($p < 0.05$) in the model, in order, accuracy in the voiceless environment (65.5 percent) was significantly lower than the flap environment (the reference level, 70.4 percent), while accuracy in the voiced environment (75.7 percent) was significantly higher than the flap environment. In addition, participants performed significantly better at discriminating variants of the /aj/ diphthong (77.1 percent) than the /aw/ diphthong (63.9 percent), and performance was better in-lab (72.7 percent) than online (69.0 percent). Importantly, there was a significant interaction between diphthong and voiceless environment, due to the flap environment having a 9.4 percentage point accuracy advantage over the voiceless environment in the /aj/ diphthong but only a 0.3 percentage point accuracy advantage over

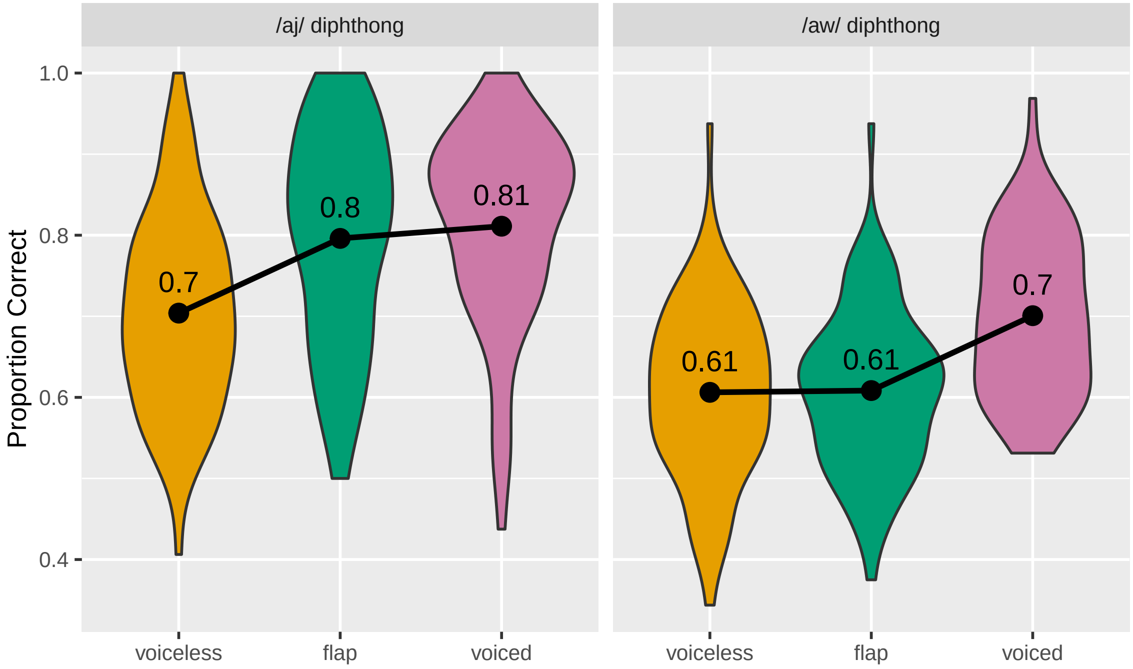

Figure 3.3: Discrimination of diphthong variants in Experiment 1

the voiceless environment in the /aw/ diphthong. According to post-hoc paired *t*-tests, the flap discrimination advantage is significant for the /aj/ diphthong (t_{78} = 6.82, $p < 0.01$) but not for the /aw/ diphthong (t_{78} = 0.1698, $p = 0.87$). There was also a significant interaction between diphthong and voiced environment, due to the voiced environment having a discrimination advantage over the flap environment of 1.4 percentage points in the /aj/ (not significant: t_{78} = 1.31, $p = 0.20$) diphthong but 9.2 percentage points in the /aw/ diphthong (significant: t_{78} = 7.79, $p < 0.01$). Finally, in this model there was an interaction between diphthong and medium. The performance advantage in the lab-based experiments over the online experiments was larger for the /aj/ diphthong (4.5 percentage points) than the /aw/ diphthong (2.8 percentage points).

3.1.3.2 Items

There was a relatively small number of items in this experiment, due to the small number of minimal pairs like *writing/riding* that exist (as well as the need to match them in the voiceless and voiced environments with additional real words that share the same onset). To illustrate the by-item findings, Table 3.4 shows the "voiced advantage" and "flap advantage"—the discrimination advantage of those environments relative to the voiceless environment—separately for each set of items sharing the same onset (and thus the same

Table 3.3: Mixed effects logistic regression model for Experiment 1

term	estimate	std.error	statistic	p.value
(Intercept)	0.961	0.058	16.587	0.000
environmentvoiceless	-0.279	0.053	-5.222	0.000
environmentvoiced	0.257	0.050	5.109	0.000
diphthongaw	-0.698	0.072	-9.657	0.000
mediumonline	-0.218	0.098	-2.225	0.026
environmentvoiceless:diphthongaw	0.522	0.091	5.733	0.000
environmentvoiced:diphthongaw	0.366	0.097	3.791	0.000
environmentvoiceless:mediumonline	0.035	0.106	0.334	0.738
environmentvoiced:mediumonline	0.049	0.099	0.490	0.624
diphthongaw:mediumonline	0.149	0.076	1.960	0.050
environmentvoiceless:diphthongaw:mediumonline	-0.028	0.182	-0.155	0.877
environmentvoiced:diphthongaw:mediumonline	-0.285	0.193	-1.479	0.139

Table 3.4: Item differences in Experiment 1 (advantages measured as percentage point difference)

onset	Voiced advantage	Flap advantage
r	10.8	12.2
s	10.8	6.6
kl	7.0	1.7
d	12.0	-1.2

vowel tokens). The flap environment was the reference level in the model because the initial hypothesis predicted a flap advantage over the other two conditions. Because the results instead found both a flap advantage and a voiced advantage over the voiceless environment, here in the by-items result breakdown the voiceless environment is the "reference level". All word sets show a voiced advantage of more than 7 percentage points, while only the word sets with an /aj/ diphthong (onsets /ɹ/ and /s/) have a flap advantage comparable in size.

3.1.4 Discussion

3.1.4.1 Discrimination and Partial Contrast

The initial prediction, following Whalen et al. (1997), Boomershine et al. (2008), Celata (2008), and Barrios et al. (2016), was that discrimination of raised and non-raised diphthong variants would be better in the environment where they can create different words (before a flap) than in the environments where they cannot (before a voiceless or voiced sound). In the results of the experiment, there was an advantage found for the flap environment but it was only an advantage compared to the voiceless environment (not the voiced one), and it only applied in the /aj/ diphthong and not in the /aw/ diphthong.

The unexpected finding of high discrimination accuracy in the voiced environment has a few possible causes. First, it could be related to the significantly greater vowel length in the voiced environment, giving participants more time to evaluate the vowels, leading to better discrimination. Second, it is also possible that this is a result of a certain asymmetry between the voiceless and voiced environments. The unexpected overapplication of Canadian Raising in the voiced context (e.g., [ˈɹʌjd] *ride*) is perhaps more surprising to a Canadian listener than the unexpected underapplication of Canadian Raising in the voiceless context (e.g., [ˈɹajt] *write*), because although both violate the rules of Canadian English, pronunciations without Canadian Raising are more familiar from other dialects of English.[3] In addition, an underapplication of Canadian Raising might also be less surprising or noticeable to a listener due to it being the underlying or canonical form of the vowel. These special characteristics of the voiced environment compared to the voiceless environment (greater length and perhaps greater surprisal for one of the diphthong variants) make the comparison between the voiceless and flap environments likely more interesting for understanding the effect of partial contrast on perception.

The unexpected isolation of the flap advantage to the /aj/ diphthong (with no advantage in the /aw/ diphthong) has a few possible causes. One explanation is that this difference arose because raised and non-raised diphthongs more clearly created different words in the flap condition for /aj/ than for /aw/. The *writing/riding* and *sighting/siding* distinctions use relatively common and recognizable words, while the *clouting/clouding* and *doubting/Dowding* distinctions rely on some uncommon or obscure words (*clouting* and *Dowding*). It is possible that participants perceived themselves to be discriminating different words in the flap condition for /aj/ but not for /aw/. This lexical explanation for the unexpected diphthong difference would mean that partial contrast does lead to a discrimination advantage, but the key to this discrimination advantage (in the case of raised and non-raised diphthongs) is not just a phonological environment where they *could* create different words (i.e., before a flap) but rather a lexical environment where they *do* create different words (e.g., in *writing/riding*).

Another explanation is that the discrimination advantage in the flap environment occurring for /aj/ but

[3] That might be the case based on other dialects, but based on patterns in Canadian English itself, it does not appear that an unexpected overapplication of Canadian Raising would be more surprising than an unexpected underapplication. While Hall (2005) finds underapplication of Canadian Raising to be more common than overapplication in a small sample of speakers from Ontario, Fullerton (2019) instead finds *overapplication* to be more common in a larger sample of speakers from Ontario and British Columbia.

not for /aw/ is related not to the items that were available and chosen but to the diphthongs themselves, whether on an acoustic level or a phonological level. There are various acoustic differences between the diphthongs, most obviously that /aj/ ends in a front vowel/glide while /aw/ ends in a back vowel/glide; additionally, there is a greater difference between raised and non-raised diphthongs for /aj/ than for /aw/ (as found by the phonetic analysis in Chapter 2 for the experimental stimuli in this dissertation, but also found for a broader sample of speakers of Canadian English in Hall, 2015). These acoustic differences between /aj/ and /aw/ could help explain other perceptual findings (like the overall higher accuracy for discrimination of [ʌj]~[aj] than [ʌw]~[aw]) but it is unclear how these acoustic differences could explain why the context effect (advantage for the flap environment over voiceless environment) was found for /aj/ but not for /aw/. Turning to phonological differences between the diphthongs that could help explain the diphthong discrepancy in the context effect, it could be that the two diphthongs have a different status in the phonology of English (or in the phonology of the relevant raising dialects) such that only the /aj/ variants are marked as or recognized by listeners as contrastive (capable of creating different words) in the flap environment. One reason to consider this plausible is evidence from certain dialects and certain speakers of raised diphthongs becoming lexicalized in the flap environment for /aj/ but not for /aw/. For example, there is evidence of speakers that have [ʌj] in *spider* and *cider* and an [ʌj]~[aj] minimal pair in *idle/idol* (Vance, 1987; Fruehwald, 2008). Unlike the minimal pairs like *writing/riding* that have received attention in this dissertation, these raised diphthongs are not phonologically predictable. If [ʌj]~[aj] and [ʌw]~[aw] do have a different status in the phonology of English and this explains the unexpected diphthong difference then it would mean that partial contrast leads to a discrimination advantage, and the key is just to have a phonological environment where the segments can create different words (i.e., before a flap)—but the assumption that this applies to /aw/ in addition to /aj/ was incorrect.

These two explanations make different predictions for experiments that decouple lexicality and diphthong. In an experiment that uses non-word items (to remove lexicality) while testing both diphthongs, the phonological explanation would predict a flap advantage over the voiceless environment for /aj/ (but not /aw/), while a (purely) lexical explanation would not. Given an experiment that holds diphthong constant but manipulates lexicality in the flap condition, however, the lexical explanation would predict a discrimination difference while a (purely) phonological explanation would not. There is also, of course, the possibility that both the phonological and lexical explanations are correct. This is to say that [ʌj]~[aj] has a different status than [ʌw]~[aw] in the phonology of English, which contributed to the finding of a flap advantage for /aj/ but not for /aw/, and there was an additional discrimination advantage in the flap environment for

/aj/ as a result of the actual minimal pairs used as stimuli. The following three experiments in this dissertation were designed with the intention of distinguishing between these three possibilities—phonological explanation, lexical explanation, or both—finding evidence for both (the phonological explanation in Experiment 2, and the lexical explanation in Experiments 3 and 4).

3.1.4.2 Other Findings

There was also a finding of better overall performance for /aj/ than /aw/ across phonological environments, which is expected based on the finding in the phonetic analysis in Chapter 2 that there is a larger phonetic difference between [ʌj]~[aj] than between [ʌw]~[aw]. This might also be related to frequency, as /aj/ is a more common vowel in general; 3.6 times more common in type frequency and 4.1 times more common in token frequency, according to the SUBTLEX corpus (Brysbaert and New, 2009) accessed through IPhOD database (Vaden et al., 2009). This too could be relevant as a potential factor behind [ʌj] lexicalizing in some dialects and possibly emerging as its own phoneme (Vance, 1987; Fruehwald, 2008). Finally, the /aj/ tokens had somewhat longer durations (29 ms average difference across the three contexts) than the /aw/ tokens, which could also contribute to the overall higher discrimination for [ʌj]~[aj] than [ʌw]~[aw] in Experiment 1.

Turning to the differences between the lab and online, overall accuracy was somewhat higher in the lab than online (with the difference being a little bigger for the /aj/ diphthong than the /aw/ diphthong). The lower accuracy from the online participants does not necessarily indicate that they were less attentive or that their data quality is lower; it could also reflect the linguistics training of the in-lab participants. Other than this, the patterns in the data were the same in both settings. This strengthens the generalizability of the findings of this experiment, and it suggests that these two methods of data collection are comparable for the purposes of this dissertation. With online experiments, a researcher might worry about their lack of control over the equipment and environment (as well as participants possibly having their attention divided between the experiment and something else). On the other hand, with lab-based experiments, a researcher might worry that the population tested (primarily young and educated, with a disproportionate chance of having taken courses in the field being studied, whether linguistics, psychology, etc.) would not be representative of the broader population. While these remain valid concerns, it appears that none of these factors mattered sufficiently for this experiment, given that the main patterns in the results were largely the same in-lab and online. Given this, and past findings on the legitimacy of online data collection (Gosling et al., 2004; Sprouse, 2011; Hauser and Schwarz, 2016), the rest of the experiments in this dissertation will be online.

3.2 Experiment 2

3.2.1 Rationale

This experiment tests the ability of speakers of Canadian English to discriminate raised and non-raised diphthongs ([ʌj]~[aj] and [ʌw]~[aw]) in the same three phonological environments as Experiment 1 (before a voiceless /t/, a voiced /d/, and a flap [ɾ]) but in non-words instead of real words. Thus, compared to Experiment 1, this experiment eliminates lexicality to test the effect of the diphthongs themselves. The primary hypothesis, in line with findings in Whalen et al. (1997), Boomershine et al. (2008), Celata (2008), Barrios et al. (2016), and Experiment 1, is that discrimination of the vowel variants will be better in the flap environment than the voiceless environment, but only for the /aj/ diphthong (and not the /aw/ diphthong). If this hypothesis holds, it would provide evidence for the phonology-based explanation mentioned in the discussion of Experiment 1, which says that [ʌj]~[aj] and [ʌw]~[aw] have a different status in the phonology of English and that this is the reason (or at least one reason) for the flap advantage (over the voiceless environment) found in Experiment 1 for /aj/ but not /aw/. Secondary predictions, based on Experiment 1, are better overall discrimination in /aj/ than /aw/, and better discrimination for the voiced environment than the voiceless environment.

3.2.2 Method

3.2.2.1 Participants

Forty-nine native speakers of English (born and currently living in Canada, mean age = 29.2, SD = 7.9, 27 men and 22 women) were recruited to participate in a browser-based online experiment through Prolific (an additional two participants were tested but excluded from analysis for reasons of language background or country status). The geographic distribution of online participants is provided in Figure 3.4. Most participants indicated little-to-no experience with linguistics. Participants provided informed consent using an online consent form. Participants were paid CAD$4.00 for a session of approximately 15 minutes.

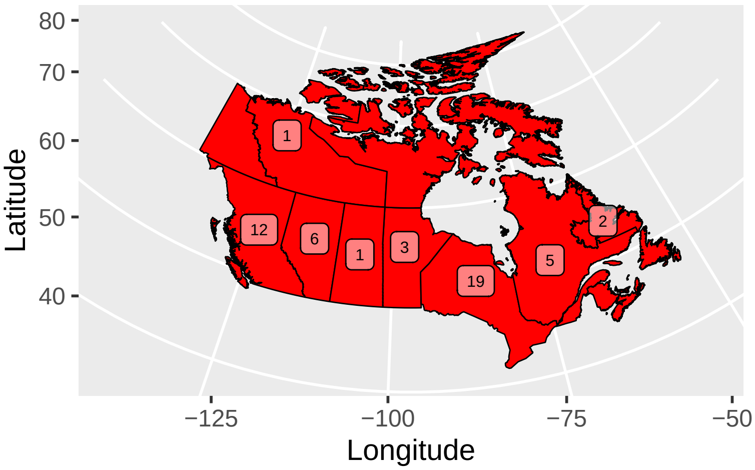

Figure 3.4: Distribution of Participants in Experiment 2 (province of birth)

Table 3.5: Experiment 2 stimuli

Voiceless	Flap	Voiced
fidight	fidighting	fidide
kuvight	kuvighting	kuvide
stazight	stazighting	stazide
fidaut	fidauting	fidaud
kuvaut	kuvauting	kuvaud
stazaut	stazauting	stazaud

3.2.2.2 Items

The stimuli in this experiment are the 18 non-words with an /aj/ or /aw/ diphthong in Table 3.5. These are based on bisyllabic "roots" that have stress on the second syllable, chosen by taking monosyllabic real words with an /aj/ or /aw/ diphthong before a /t/ (e.g., *tight*), changing the onset to make a non-word (except in one case), and then adding an additional prefix to make it more obviously a non-word (this process can be seen in, e.g., *tight* ⟶ *dight* ⟶ *fidight*). Then, voiced and flap versions were created as well (*fidide* and *fidighting*). Stimulus creation in Experiment 2 was identical to Experiment 1, using the same two speakers and the same process of splicing (including duration manipulation).[4] The vowel durations in the different environments are in Table 3.6.[5]

[4] The one difference is that only raised versions of the flap words (like *kuvighting*) were recorded, compared to Experiment 1, where both *writing* and *riding* were recorded and used as sources for the flap environment stimuli. This could have a small effect on vowel duration but not vowel quality because the vowels were still sourced from the voiceless and voiced environment words (for raised and non-raised vowels respectively).

[5] Broken down by diphthong rather than gender, there was a very slight pattern of /aj/ diphthongs being longer in duration than /aw/. In Experiment 1 the /aj/ tokens were on average 29 ms longer, while here in Experiment 2 they are on average 5 ms longer (214 ms versus 221 ms in the voiceless context for /aj/ and /aw/, respectively; 200 ms versus 196 ms in the flap context,

Table 3.6: Experiment 2 stimulus vowel durations (ms)

Gender	Voiceless	Voiced	Flap
Male	201.4	326.7	191.0
Female	233.9	365.6	204.5
Average	217.6	346.2	197.8

The stimuli were normalized to an average root mean square intensity of 70 dB Sound Pressure Level (SPL). The final result was 144 sound files: 18 words × two speakers × two variants of the diphthong (raised and non-raised) × two versions.

3.2.2.3 Procedure

This experiment was an AXB discrimination task mirroring the online version of Experiment 1. There were 144 trials, with 48 from each environment (voiceless, flap, and voiced) interspersed.

3.2.3 Results

3.2.3.1 Mixed Effects Model

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 3.5. In contrast to Experiment 1, which had a 9.4 percentage point flap advantage (over the voiceless condition) in the /aj/ diphthong and a 0.3 percentage point advantage in the /aw/ diphthong, in Experiment 2 there was a 3.4 percentage point flap advantage in /aj/ and a 2.1 percentage point flap *disadvantage* in /aw/.

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The fixed effects were environment (three levels: voiceless, voiced, and flap) and diphthong (two levels: aj and aw). The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level for environment was "voiceless" (because the primarily hypothesis for Experiment 2 is a flap advantage over the voiceless environment only for the /aj/ diphthong) and the reference level for diphthong was "aj". The approach to

and 355 versus 337 ms in the voiced context).

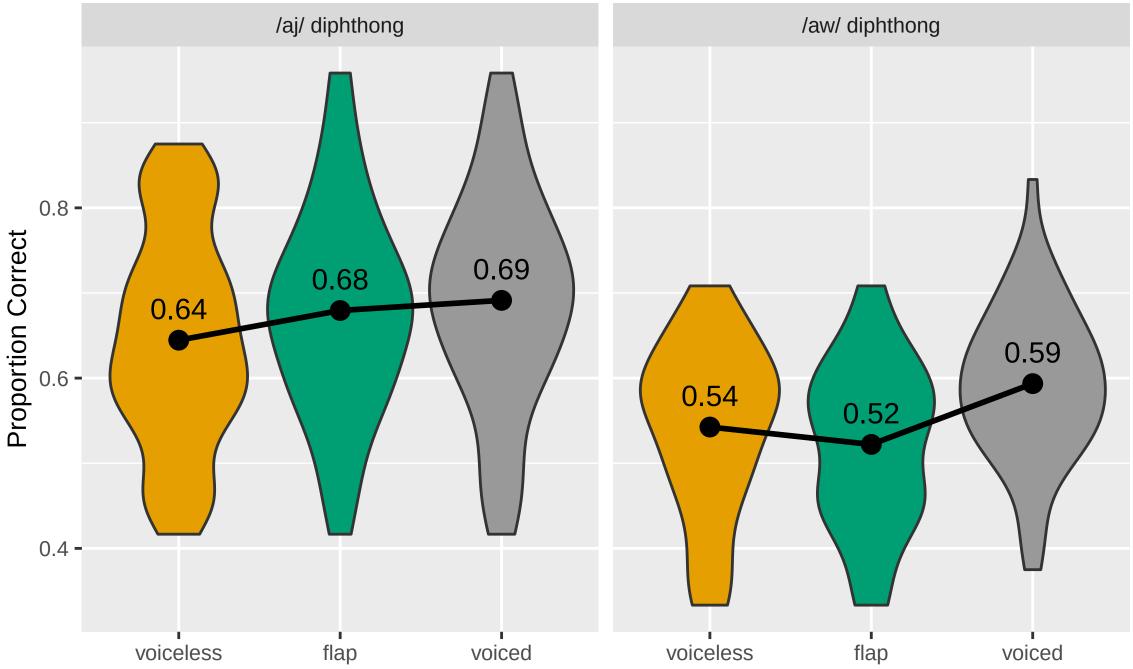

Figure 3.5: Discrimination of diphthong variants in Experiment 2

Table 3.7: Mixed effects logistic regression model for Experiment 2

term	estimate	std.error	statistic	p.value
(Intercept)	0.471	0.039	11.965	0.000
environmentflap	0.038	0.063	0.608	0.543
environmentvoiced	0.217	0.075	2.874	0.004
diphthongaw	-0.511	0.050	-10.275	0.000
environmentflap:diphthongaw	-0.241	0.121	-1.995	0.046
environmentvoiced:diphthongaw	-0.006	0.122	-0.053	0.958

random effects was to use the maximal structure justified by the experimental design that does not result in a failure to converge or a singular fit (Barr et al., 2013). The result was by-subjects and by-items random intercepts, as well as by-subjects and by-items random slopes for environment. The items were the rows in Table 3.5, which shared the same onset and the same vowel tokens. The results of the mixed effects analysis are presented in Figure 3.7.

To summarize the significant findings ($p < 0.05$) in the model, in order, overall accuracy in the voiced environment (64.2 percent) was significantly higher than in the voiceless environment (59.4 percent) and overall accuracy was lower for the /aw/ diphthong (55.3 percent) than the /aj/ diphthong (67.2 percent). Importantly, as in Experiment 1, there was a significant interaction between diphthong and voiceless environment, due to the flap environment having a 3.4 percentage point advantage over the voiceless

Table 3.8: Item differences in Experiment 2 (advantages measured as percentage point difference)

diphthong	onset	Voiced advantage	Flap advantage
aj	f	12.5	2.0
aj	k	3.8	6.1
aj	s	-2.3	2.3
aw	f	3.8	-3.1
aw	k	4.3	-3.1
aw	s	7.1	0.0

environment in the /aj/ diphthong but a 2.1 percentage point *disadvantage* in the /aw/ diphthong. However, in a post-hoc paired *t*-test, the flap advantage in the /aj/ diphthong was not quite significant (t_{48} = 1.67, *p* = 0.10), which differs from Experiment 1.

3.2.3.2 Items

To illustrate the by-item findings, Table 3.8 shows the voiced advantage and flap advantage (calculated with reference to the voiceless environment) separately for each series of items sharing the same onset and diphthong (and thus sharing the same vowel tokens). All but one series exhibited a positive voiced advantage, and all three /aj/ series exhibited a positive flap advantage, although they were mostly small in size.

3.2.4 Discussion

3.2.4.1 Discrimination and Partial Contrast

As in Experiment 1, the relation in performance between the flap and voiceless environments differed between diphthongs to a statistically significant degree, with there being more of a flap advantage in /aj/ than /aw/. Unlike Experiment 1, however, the flap advantage in /aj/ was itself not statistically significant in a post-hoc *t*-test. This discrepancy can be attributed to the finding of a numerically small (2.3 percentage point) flap disadvantage in the /aw/ condition in Experiment 2. The /aj/ flap advantage in this experiment was 3.4 percentage points (and not statistically significant) if measured against the /aj/ voiceless environment, but 5.5 percentage points (and statistically significant, based on the significant interaction) if measured against the /aj/ voiceless environment *and then compared* with the corresponding relation between the flap and voiceless environments in the /aw/ diphthong.

Although this experiment does not provide (statistically significant) evidence that a flap advantage, as defined in Experiment 1, exists for the /aj/ diphthong in the absence of lexicality, it does provide evidence that the two diphthongs differ in the relation between their flap and voiceless environments in the absence of lexicality. To return to the main question remaining from Experiment 1 (why there was a flap advantage found for /aj/ but not /aw/), the results of this experiment suggest that the finding in Experiment 1 was not purely related to lexicality and the items used (the fact that the stimuli in the flap condition for /aj/ were good minimal pairs like *writing/riding* while the flap condition for /aw/ included more dubious minimal pairs like *clouting/clouding*). Instead, there appears to be some factor related to the diphthongs themselves. As it seems unlikely that there is an acoustic difference between these diphthongs that would cause them to be affected differently by phonological environment, this was proposed as the phonological explantion—/aj/ and /aw/ have a different status in the phonology of English, or at least the phonology of these relevant raising dialects ([ʌj]~[aj] is marked or recognized as contrastive in the flap environment, while [ʌw]~[aw] is not). The findings of Experiment 2 support this phonological explanation, and they compare with the previous findings from Western Tuscan of faster and more accurate identification of a partial contrast in the contrastive environment than the non-contrastive environment (Celata, 2008).

While these results support the phonological explanation, they do not rule out the lexical explanation as an additional factor. The flap advantage found for the /aj/ but not /aw/ diphthong in Experiment 1 could have been a result of the diphthongs themselves *and* the items used (the better minimal pairs for /aj/). In fact, this possibility is suggested by the fact that this experiment failed to completely replicate Experiment 1 with a clear flap advantage over the voiceless environment for /aj/.

3.2.4.2 Other Findings

Experiment 2 also replicated the finding in Experiment 1 of more accurate discrimination in the voiced environment, and better overall performance for /aj/ than /aw/ across phonological environments. The continued finding of the diphthong advantage provides increasing evidence that listeners simply have an easier time discrimination [ʌj]~[aj] than [ʌw]~[aw], which could (alongside the finding that the difference is especially big in the flap condition) help explain [ʌj] lexicalizing in some dialects and possibly emerging as its own phoneme. One difference from Experiment 1 is that /aj/ tokens were somewhat longer than /aw/ tokens in Experiment 1 (29 ms), which might have explained some of the overall discrimination advantage for /aj/ over /aw/, but in Experiment 2 the durational difference between /aj/ and /aw/ was much smaller

(5 ms in the same direction), meaning that durational differences are less plausibly a component of the overall discrimination advantage for /aj/ in Experiment 2.

Chapter 4

Manipulating Lexicality

This chapter presents two experiments building on the basic findings about the context-based effect of partial contrast on discrimination from Experiments 1 and 2. Experiment 1 found better discrimination of raised and non-raised diphthongs in the flap (contrastive) environment than the voiceless (non-contrastive) environment, but only for /aj/ and not /aw/. Experiment 2 found comparable results using non-words, suggesting the diphthong difference found in Experiment 1 was in fact related to the diphthongs themselves and not just the lexical items used (the fact that the items in the flap environment created better minimal pairs for /aj/, e.g., *writing/riding*, than for /aw/, e.g., *clouting/clouding*). The experiments in this chapter return to real words to investigate the role of lexicality in the effect of partial contrast on discrimination, specifically testing whether the diphthong difference in Experiment 1 was also related to the lexical items used and not just the diphthongs. Experiment 3 tests the role of lexicality using a similar design to the first two experiments, while Experiment 4 uses a more novel design that tests discrimination of the exact same contrast (*writing/riding*) in different environments that are meant to allow both interpretations or encourage one interpretation. Both find evidence that lexicality (i.e., whether or not the items being discriminated are minimal pairs) plays a role in discrimination. These findings suggest that the unexpected diphthong differences in Experiment 1 were indeed also related to the lexical items used, and they illuminate another component of the effect of partial contrast on discrimination: lexicality.

4.1 Experiment 3

4.1.1 Rationale

This experiment tests the ability of speakers of Canadian English to discriminate raised and non-raised diphthongs in real words with the /aj/ diphthong that have been selected such that the diphthong variants [ʌj]~[aj] either cause a minimal pair (like *writer/rider*) or do not (like *fighter*). Thus, compared to Experiment 1 (which confounded diphthong and lexicality) and Experiment 2 (which discarded lexicality to isolate the effect of diphthong), Experiment 3 holds diphthong constant to isolate the effect of lexicality. The primary hypothesis, based on findings in Experiments 1 and 2, is that the relation between the voiceless and flap environments will differ between those two sets of words. More specifically, *writer/rider* (flap environment with a lexical distinction) will have a larger advantage over *write* (its voiceless environment counterpart) than *fighter* (flap environment without a lexical distinction) will have over *fight* (its voiceless counterpart). A secondary prediction is better discrimination for the voiced environment than the voiceless environment, in line with the previous experiments.

4.1.2 Method

4.1.2.1 Participants

Experiment 3 was separated into two versions (one whose flap environment has a lexical contrast with stimuli like *writer/rider*, and one whose flap environment does not, with stimuli like *fighter*) to be run on separate participants. This decision was made due to findings (to be reported in Chapter 5 of this dissertation) that the presence or absence of experimental conditions can influence the presence or absence of the flap advantage. Forty-six speakers of English, born and currently living in Canada (mean age = 26.4, SD = 7.2, 29 men and 17 women), participated in the lexical contrast variant of this online browser-based experiment (an additional three participated but their data was excluded for linguistic/demographic reasons). For the no lexical contrast variant of the experiment, there were 45 participants (mean age = 27.2, SD = 6.4, 21 men, 23 women, 1 other), with one additional participant excluded. The geographic distribution of participants across both versions of the experiment is provided in Figure 4.1. Participants were paid CAD$2.00 for an experiment lasting approximately 10 minutes.

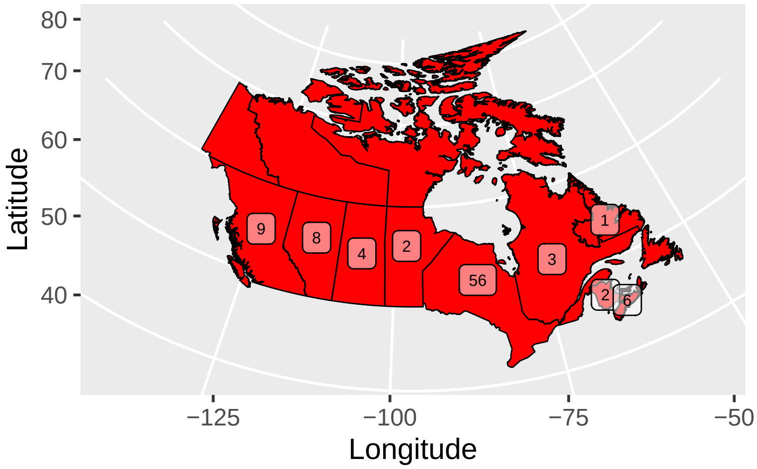

Figure 4.1: Distribution of Participants across both versions of Experiment 3 (province of birth)

Table 4.1: Experiment 3 stimuli

Group	Voiceless	Flap	Voiced
1	write	writer (rider)	ride
1	tight	title (tidal)	tide
2	light	lighter	lied
2	fight	fighter	defied

4.1.2.2 Items

The stimuli in this experiment are the 12 words (or 14 words if counting *writer/rider* and *title/tidal* as each being two words) with an /aj/ diphthong in Table 4.1. The first group (the first two rows) has a lexical distinction in the flap environment, intended to correspond to the /aj/ stimuli in Experiment 1. The second group (the third and fourth rows) lacks a lexical distinction in the flap environment, and is intended to correspond to the /aw/ stimuli in Experiment 1 (which did not have clear minimal pairs in the flap environment). These are the portions of the present experiment that were run on different participants.

Stimulus creation in Experiment 3 mirrored the process used in Experiments 1 and 2 (using the same two speakers), with the exception of duration manipulation. For all environments in all of these first three experiments, the raised and non-raised vowels were modified to match each other in duration within each environment. For example, to create the stimuli in the voiceless environment (in a word like *right*), a raised vowel was taken from a recording of *right*, a non-raised vowel was taken from a recording of *ride*, and then these were both spliced into a different recording of *right*, taking on the duration of that original vowel in that second recording of *right*. In Experiments 1 and 2, this process was carried out in the same way for all three environments. However, in Experiment 3, the duration in the flap environment was modified to match the duration in the voiceless environment. This change was done to make the discrimination results (a small amount) more comparable, at the cost of (a small amount of) naturalness. The duration of the vowels in the voiced environment in Experiment 3 were not changed to match the voiceless environment in this way, because it would be a much bigger change (possibly affecting perception of voicing of the consonant) and because the voiced condition is of less theoretical interest.

The stimuli were normalized to an average root mean square intensity of 70 dB Sound Pressure Level (SPL). The final result was 96 sound files: 12 words × two speakers × two variants of the diphthong (raised and non-raised) × two versions.

4.1.2.3 Procedure

This experiment was an AXB discrimination task mirroring the online version of Experiment 1. In each version of the experiment (group 1 and 2 in Table 4.1) there were 96 trials, with 32 from each environment (voiceless, flap, and voiced) interspersed.

4.1.3 Results

4.1.3.1 Mixed Effects Model

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 4.2. In contrast to Experiment 1, which had a 9.4 percentage point flap advantage in the /aj/ diphthong and a 0.3 percentage point advantage in the /aw/ diphthong, Experiment 3 had a 1.5 percentage point flap advantage in the lexical distinction sub-experiment (intended to correspond to the lexical distinction in the /aj/ stimuli in Experiment 1) and a 6.9 percentage point flap *disadvantage* in the no lexical distinction sub-experiment (intended to correspond to the lack of lexical distinction in the /aw/ stimuli in Experiment 1).

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The fixed effects were environment (three levels: voiceless, voiced, and flap) and sub-experiment (two levels: lexical distinction and no lexical distinction). The contrast coding used for these categorical

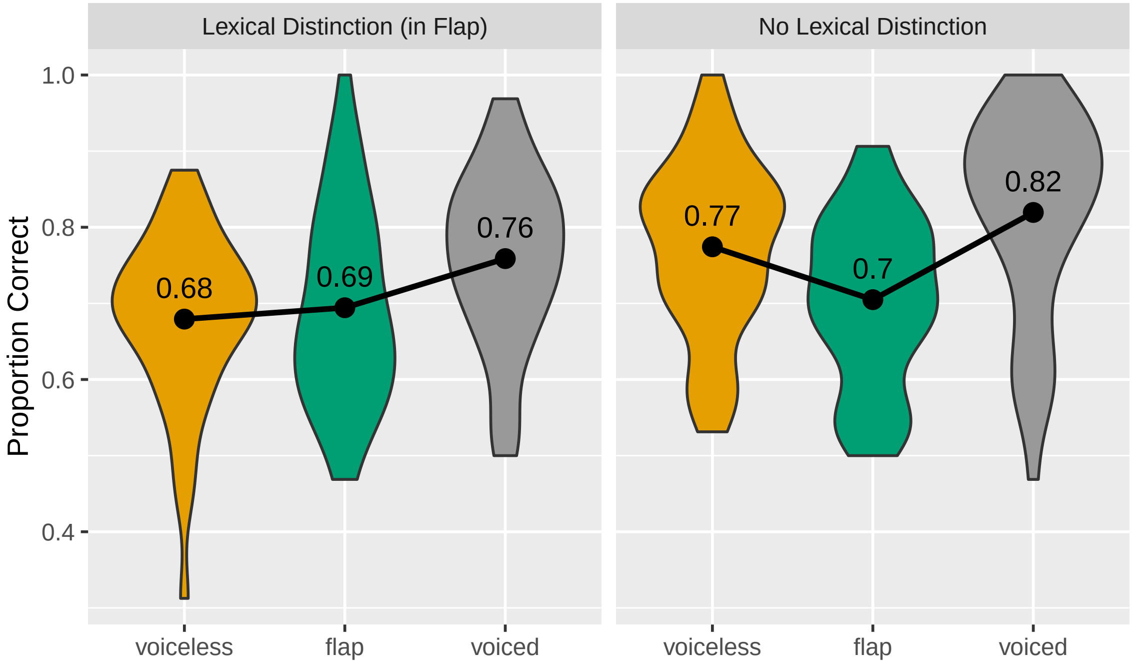

Figure 4.2: Discrimination of diphthong variants in Experiment 3

Table 4.2: Mixed effects logistic regression model for Experiment 3 ('subexplexdis' refers to the lexical distinction versus no lexical distinction sub-experiments)

term	estimate	std.error	statistic	p.value
(Intercept)	1.127	0.136	8.253	0.000
environmentflap	-0.156	0.060	-2.610	0.009
environmentvoiced	0.355	0.064	5.572	0.000
subexplexdis	-0.340	0.273	-1.248	0.212
environmentflap:subexplexdis	0.459	0.120	3.828	0.000
environmentvoiced:subexplexdis	0.119	0.127	0.935	0.350

variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level for environment was "voiceless" (because the primarily hypothesis for Experiment 3 is a flap advantage over the voiceless environment only for the lexical distinction sub-experiment) and the reference level for sub-experiment was the no lexical distinction sub-experiment. The approach to random effects was to use the maximal structure justified by the experimental design that does not result in a failure to converge or a singular fit (Barr et al., 2013). The result was by-subjects and by-items random intercepts, with no random slopes. The items were the rows in Table 4.1, which shared the same onset and the same vowel tokens. The results of the mixed effects analysis are presented in Figure 4.2.

To summarize the significant findings ($p < 0.05$) in the model, in order, overall accuracy in the flap environ-

Table 4.3: Item differences in Experiment 3 (advantages measured as percentage point difference)

subexp	onset	Voiced advantage	Flap advantage
nolexdis	f	8.9	-11.4
nolexdis	l	0.1	-2.5
lexdis	r	-0.1	-1.4
lexdis	t	16.0	4.3

ment (70.0 percent) was significantly lower than in the voiceless environment (72.6 percent), but overall accuracy in the voiced environment (78.9 percent) was higher than the voiceless environment. In addition, overall accuracy was also lower in the lexical distinction sub-experiment (71.1 percent) than the no lexical distinction sub-experiment (76.6 percent). Importantly, there was a significant interaction between sub-experiment and the flap environment, due to the flap environment having a 1.5 percentage point advantage over the voiceless environment in the lexical distinction sub-experiment, but a 6.9 percentage point disadvantage compared to the voiceless environment in the no lexical distinction sub-experiment. In a post-hoc paired t-test, the flap advantage in the lexical distinction sub-experiment was not statistically significant ($t_{45} = 0.93$, $p = 0.36$), which differs from Experiment 1.

4.1.3.2 Items

To illustrate the by-item findings, Table 4.3 shows the voiced advantage and flap advantage (calculated with reference to the voiceless environment) separately for each series of items sharing the same onset (and thus sharing the same vowel tokens). Two of the four series had a voiced advantage near zero, which is a less consistent voiced advantage than was found in the previous two experiments. For the flap advantage, three of the four series had a negative flap advantage (i.e., a flap disadvantage), but there was a split between the lexical distinction and no lexical distinction sub-experiments, with the latter being more clearly into the realm of a flap disadvantage.

4.1.4 Discussion

4.1.4.1 Discrimination and Partial Contrast

The patterns found in Experiment 3 were similar to those found in Experiment 2, although the theoretical implications are different. Both experiments saw a weaker flap advantage than Experiment 1; in fact, for both Experiments 2 and 3 there was no significant difference between the flap and voiceless environments in the half of the experiment where a flap advantage was expected (the /aj/ non-words in Experiment 2, and the /aj/ words in the lexical distinction sub-experiment in Experiment 3). However, in both experiments the small flap advantage in one half was statistically significant when compared against the flap disadvantage in the other half (the /aw/ non-words in Experiment 2, and the /aj/ words in the no lexical distinction sub-experiment in Experiment 3).

This means that neither experiment provided evidence for a flap advantage in the same way as Experiment 1, but both experiments provided evidence for factors that influence the relation between the voiceless and flap environments (regarding the accuracy of discriminating raised and non-raised diphthongs in these environments). Experiment 2 found evidence for the diphthong itself having an impact on the relation between the voiceless and flap environments (independent of the lexicality of the items), while Experiment 3 found evidence for the lexicality of the items—do raised and non-raised vowels in the flap environment create different words or not?—also affecting the relation between the voiceless and flap environments.

To return to Experiment 1, where a flap advantage (over the voiceless environment) was found for /aj/ but not /aw/ and it was unclear whether this was related to the items used or to the diphthongs themselves, the two follow-up experiments suggest that both the lexical explanation and the phonological explanation (so-called because an inherent difference between /aj/ and /aw/ regarding the relation between the voiceless and flap environments seems unlikely to be related to the acoustics of the vowels) are correct.

Given the apparent effect of both lexicality and diphthong on the flap advantage, the lack of significant difference between the voiceless and flap conditions for /aj/ non-words in Experiment 2 is more easily explained than the lack of significant difference between the voiceless and flap conditions in the lexical distinction sub-experiment in Experiment 3. In Experiment 2, the flap advantage was being driven solely by the diphthong itself, without the additional boost to discrimination of the lexical distinction. Experiment 3, however, had both an /aj/ diphthong and lexical distinction; despite this, the flap advantage was still small (1.5 percentage points) and not statistically significant. One possibility is that the lexical distinctions in Experiment 3 (*writer/rider* and *title/tidal*) were weaker or less obvious to listeners than those in Experiment 1 (*writing/riding* and *sighting/siding*), but *writer/rider* and *writing/riding* are very similar distinctions, differing mainly in suffix and syntactic function, and yet the former had a flap advantage of -1.4 percentage points and the latter a flap advantage of 12.2 percentage points. In comparison, *title/tidal* and *sighting/siding* (which at least are not morphologically related to each other) had more comparable flap advantages at 4.3 percentage points and 6.6 percentage points, respectively. In addition, it has to also be considered that whatever factor resulted in unexpectedly low performance in the flap condition in the lexical distinction sub-experiment also resulted in unexpectedly low performance in the flap condition in the no lexical distinction sub-experiment (with *fighter* and *lighter*)—indeed, that is the reason why despite these differences, a comparable interaction to Experiment 1 was also found in Experiment 3 (more of a flap advantage in the lexical distinction sub-experiment). Another salient difference between Experiments 1 and 3 is that Experiment 3 was split up into two parts to be run on two separate groups of participants. This was done due to findings (to be reported in Chapter 5) that the presence or absence of experimental conditions can influence the presence or absence of the flap advantage. The concern was that exposure to the no lexical distinction flap condition might reduce the flap advantage in the lexical distinction condition. However, it is possible that splitting up the experiment into two parts actually had the effect of reducing the flap advantage in both of those parts. Splitting up the experiment resulted in each participant having less exposure to the speakers' voices over the course of their participation (96 trials instead of the 192 in Experiment 1), and (as will be discussed in Chapter 5) there is reason to believe that the flap advantage is dependent on exposure to the speakers' voices.

Another way to look at this finding is to say, setting aside the voiced condition, all of the other results are around 70 percent, except for the voiceless condition in the no lexical distinction sub-experiment being at 77.4 percent. Perhaps there is something special about those items (*light* and *fight*). However, it is again not clear what that would be, and this line of thought relies on directly comparing values between the two sub-experiments, which might be misleading because they use different vowel tokens (like their equivalents in the previous experiments, which were different diphthongs) and they were run on different participants (unlike in the past two experiments). Because of these two factors, the intended analysis for this experiment was to compare the patterns between the two sub-experiments (particularly the difference between voiceless and flap environments in one compared to the other), rather than to directly compare, for example, the flap environment in one sub-experiment with the flap-environment in the other.

While the unexpectedly low performance in the flap environments remains an open question, the previously discussed main finding of this experiment (that the lexicality of the items appears to affect the relation between the voiceless and flap environments) does help us better understand the results of Experiment 1 and the effect of partial contrast on discrimination. The next experiment uses a different design to further investigate the effect of lexicality on discrimination in the context of partial contrast.

4.1.4.2 Other Findings

As in both of the past experiments, the voiced environment again showed a notably high rate of accuracy for discriminating diphthong variants. The most likely explanation for this is the longer vowel duration in this environment, providing the listener with a longer window to evaluate the vowels—or the fact that an overapplication of Canadian Raising in this context (e.g., [ˈɹʌjd] *ride*) is more surprising than the underapplication of Canadian Raising in the voiceless environment (e.g., [ˈɹajt] *write*), because an underapplication of Canadian Raising is more familiar from other dialects and is using the underlying or canonical form of the vowel.

4.2 Experiment 4

4.2.1 Rationale

This experiment tests the ability of speakers of Canadian English to discriminate raised and non-raised diphthongs in real words with the /aj/ diphthong that have been selected such that the diphthong variants [ʌj]~[aj] either cause a lexical distinction or not—similar to Experiment 3. However, this experiment was designed with the intention of varying the lexicality while keeping the exact same acoustic tokens. To this goal, compound words or phrases were built around the *writing/riding* contrast where the preceding word either allowed both interpretations (e.g., *still* [ˈɹʌjɾɪŋ, ˈɹajɾɪŋ]) or strongly encouraged one interpretation over the other (e.g., *book* [ˈɹʌjɾɪŋ, ˈɹajɾɪŋ]). The hypothesis was that, if the preceding words were successful at constraining or preferentially activating one interpretation over the other, discrimination of the *writing/riding* distinction would be better when preceded by a word that allows both interpretations than when preceded by a word that does not.

4.2.2 Method

4.2.2.1 Participants

Fifty native speakers of English (born and currently living in Canada, mean age = 29.2, SD = 8, 25 men and 25 women) were recruited to participate in a browser-based online experiment through Prolific. The geographic distribution of participants is provided in Figure 4.3. Most participants indicated little-to-no

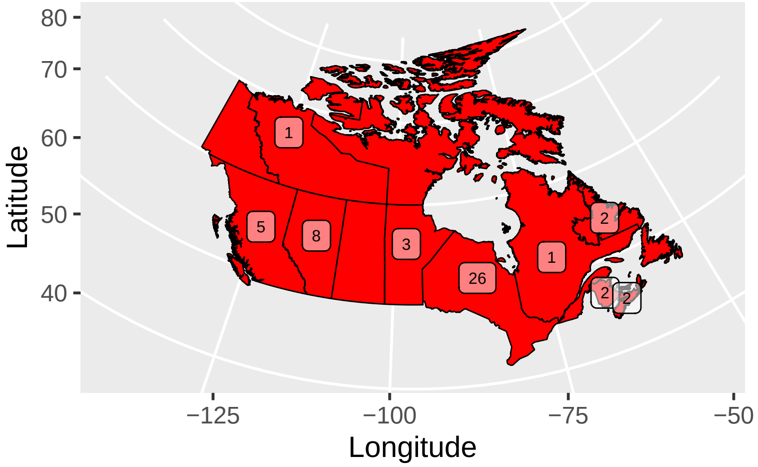

Figure 4.3: Distribution of Participants in Experiment 4 (province of birth)

Table 4.4: Experiment 4 stimuli (with COCA co-occurrences)

Preceding word	Condition	Preferred	Writing	Riding
Fiction	One-Interpretation	Writing	208	0
Book(s)	One-Interpretation	Writing	245	1
Horseback	One-Interpretation	Riding	0	624
Train(s)	One-Interpretation	Riding	0	13
Still	Two-Interpretation		157	87
Just	Two-Interpretation		172	89
Maybe	Two-Interpretation		14	4
Always	Two-Interpretation		58	19

experience with linguistics. Participants provided informed consent using an online consent form. Participants were paid CAD$2.40 for a session of approximately 10 minutes.

4.2.2.2 Items

The items in this experiment were built around the *writing/riding* contrast that was previously used in Experiment 1. Stimulus creation mirrored the process used in Experiment 1 for the *writing/riding* stimuli with the same two speakers used (new recordings were done from a session where the speakers were also recorded saying the preceding words), except that four variants of *writing* and four of *riding* were created for each speaker (instead of two of each as in Experiment 1, when *writing/riding* was only one of the items used). To create the different experimental conditions—the one-interpretation condition where either *writing* or *riding* is preferred, and the two-interpretation condition where neither interpretation is strongly preferred—the preceding words in Table 4.4 were used. These were selected primarily based on the author's judgements of semantic fit, and secondarily on co-occurrence patterns in the the Corpus of Contemporary American English (Davies, 2008), specifically how often each word appeared immediately preceding or immediately following *writing* and *riding* in the corpus. The words in the one-interpretation condition were all nouns, and in the two-interpretation condition all adverbs. The words in both conditions were evenly split between one-syllable and two-syllable. However, the stimuli in the one-interpretation condition did end up being overall 3 percent longer in duration (calculated based on the total length of all the sound files in each condition), which could possibly affect discrimination accuracy.

All eight preceding words were separately spliced onto the beginning of each speaker's recordings of *writing* and *riding*, resulting in final stimuli taking the form of *fiction-writing*, *fiction-riding*, *book-writing*, *book-riding*, *horseback-writing*, *horseback-riding*, *train-writing*, and *train-riding* in the one-interpretation condition, and *still-writing*, *still-riding*, *just-writing*, *just-riding*, *maybe-writing*, *maybe-riding*, *always-writing*, and *always-riding* in the two-interpretation condition. The intention was that an AXB trial in the one-interpretation condition would take the form of (for example) *book-writing* ... *book-writing* ... *book-riding*, and the preceding word *book* would preferentially activate the *writing* interpretation over *riding* (despite the non-raised vowel that would normally be associated with *riding*). Given the findings in Experiments 1 and 3 that suggested an effect of lexicality (raised and non-raised vowels being discriminated more accurately when they create different words than when they do not), discrimination is expected to be worse in the above example than in a trial consisting of *still-writing* ... *still-writing* ... *still-riding*, where the preceding word does not strongly point to one interpretation over the other.

It was not guaranteed that the preceding word would affect interpretation in this way. However, it is plausible in part because Hall (2005) and Fullerton (2019) find unexpected pronunciations (both underapplication and overapplication of Canadian Raising) in Canadian English. Due to this variation, some instances of *writing* probably are pronounced with non-raised vowels, and some instances of *riding* probably are pronounced with raised vowels, even if the expected prounciations are much more common. This design is a way to test discrimination of a lexical contrast with discrimination of no lexical contrast (or a weaker lexical contrast) while using the exact same acoustic tokens.[1]

In addition to these primary experimental stimuli (which had a flap environment), additional /ɹ/-onset words with a voiceless and voiced environment (the exact *right* and *ride* tokens from Experiment 1) were

[1] There did have to be a difference between the two conditions, which was the preceding word. Although the preceding words were balanced for number of syllables, the stimuli in the one-interpretation condition ended up being overall 3 percent longer in duration. This will be discussed in the results.

also presented to participants in the same format to discriminate. These environments were not intended to be part of the analysis; rather, they were included due to findings (to be reported in Chapter 5) that the flap advantage over the voiceless condition appears to rely on the presence of a voiced condition in the experiment. In the next section of this dissertation, a version of this experiment without these extra conditions will be presented as Experiment 4b.

4.2.2.3 Procedure

This experiment was an AXB discrimination task mirroring the online version of Experiment 1. There were 40 total trials in the primary experimental conditions (20 in the one-interpretation condition and another 20 in the two-interpretation condition), plus an additional 40 trials in the extra conditions that used stimuli from Experiment 1 (20 voiceless and 20 voiced). Eight of these extra condition trials were hard-coded to start at the beginning of the experiment, while the other 32 extra condition trials were interspersed with the 40 primarily experimental conditions in the rest of the experiment.

4.2.3 Results

4.2.3.1 Mixed Effects Model

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 4.4. There was a 3.4 percentage point discrimination advantage in the two-interpretation condition compared to the one-interpretation condition.

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The one fixed effect was interpretations (two levels: one and two). The contrast coding used for this categorical variable was simple coding. The reference level for interpretations was “one”. The approach to random effects was to use the maximal structure justified by the experimental design that does not result in a failure to converge or a singular fit (Barr et al., 2013). The result was by-subjects and by-items random intercepts, with by-subjects random slopes for the one predictor variable, interpretations. The items were the eight preceding words in Table 4.4. The results of the mixed effects analysis are presented in Table 4.5.

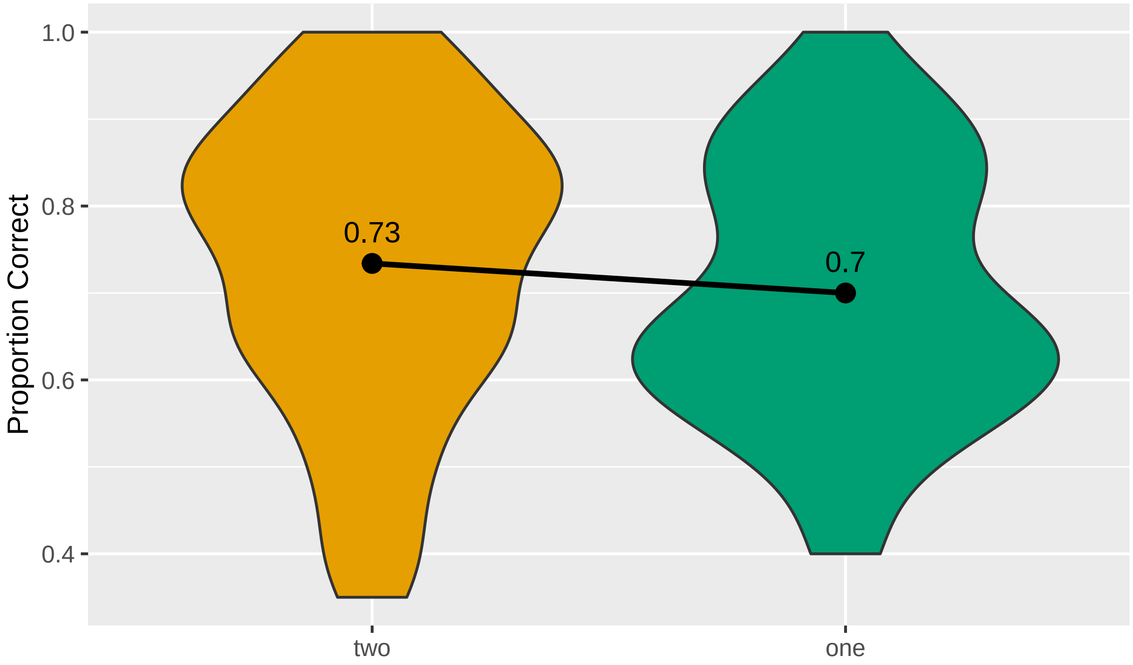

Figure 4.4: Discrimination of diphthong variants in Experiment 4

Table 4.5: Mixed effects logistic regression model for Experiment 4

term	estimate	std.error	statistic	p.value
(Intercept)	1.040	0.117	8.900	0.00
interpretationstwo	0.238	0.122	1.961	0.05

Due to the simpler experimental design, there was only one effect to be analyzed, which is the difference between one and two interpretations. Discrimination in the two-interpretation condition (73.4 percent) was significantly higher than in the one-interpretation condition (70.0 percent).

4.2.3.2 Items

The discrimination accuracies according to the preceding word are provided in Table 4.6. The two-syllable words in the one-interpretation condition (*fiction* and *horseback*) pattern noticeably lower than their one-syllable counterparts, in the range of 60–70 percent instead of 70–80 percent. There is no such difference between items in the two-interpretation condition.

Table 4.6: Item results in Experiment 4

interpretations	preword	correct
one	book	75.6
one	fiction	66.4
one	horseback	66.4
one	train	71.6
two	always	73.2
two	just	71.6
two	maybe	74.0
two	still	74.8

4.2.3.3 Other Results

In addition to the experimental items analyzed above, with the flap environment and preceding words intended to allow two interpretations or preferentially activate one interpretation, participants in this experiment were also exposed to the voiceless (*right*) and voiced environment (*ride*) items from Experiment 1. These came from different recording sessions (meaning that they use different vowel tokens than the primary experimental conditions) and they were included in the session to account for apparent findings (to be demonstrated in Chapter 5) that the flap advantage relies on the presence of the voiced condition in the experiment. As a result, they were not included in the statistical analysis and main findings. However, performance was noticeably lower in the voiceless environment (63.5 percent) and noticeably higher in the voiced environment (77.3 percent), with the two flap environments (at 70.0 and 73.4 percent) falling in the middle but closer to the voiced environment. Thus, informally, this experiment replicated the finding of a flap advantage (compared to the voiceless environment) of Experiment 1, even though the flap conditions had the preceding word, and thus they were longer in duration and more complex semantically.

4.2.4 Discussion

As hypothesized, discrimination was higher for the *writing/riding* contrast when preceded by a word that allowed both interpretations (e.g., *still-writing* ... *still-writing* ... *still-riding*) than when preceded by a word that encouraged only one interpretation (e.g., *book-writing* ... *book-writing* ... *book-riding*). This finding, together with the finding of Experiment 3, provides support for the lexical explanation of the partial contrast discrimination effects in Experiment 1, which is that one reason that a flap advantage (over the voiceless condition) was observed for /aj/ but not /aw/ is that the stimuli for /aj/ were better minimal pairs (*writing/riding* and *sighting/siding*) than the stimuli for /aw/ (*clouting/clouding* and *doubting/Dowding*). It should be noted that the result from this experiment does not necessarily uncover the size of the effect of lexicality on discrimination, because the manipulation of lexicality happened indirectly, through the preceding word. The preceding words appear to have had some success in influencing interpretation of the *writing/riding* tokens, based on the observed discrimination disadvantage for *writing/riding* when preceded by a word that allowed or facilitated only one of the two interpretations, but it still remains the case that *writing/riding* has two interpretations.

While one advantage of this design is the ability to use acoustically identical tokens for the words being discriminated (i.e., the same *writing/riding* tokens in the one-interpretation and two-interpretation conditions), there did have to be a difference between the two conditions, and that came from the preceding word. Although the preceding words were balanced for number of syllables, the stimuli in the one-interpretation condition were overall 3 percent longer in duration, which could possibly explain the lower discrimination accuracy for the one-interpretation condition. One reason to expect that this preceding word durational difference affected discrimination is the pattern in Table 4.6 where the lower accuracy for the one-interpretation condition appears to be driven by the two-syllable words in particular.

However, there are also reasons to believe that the preceding word durational difference is not the reason for the observed discrimination differences. The effect could be driven by the longer words because of longer words producing stronger lexical activation (Zhang and Samuel, 2015). Another reason to believe that these results cannot be explained by durational differences is that the voiceless tokens (*right*) added in from the Experiment 1 stimuli are 8.2 percentage points lower in accuracy than the main experimental tokens (averaged across the one-interpretation and two-interpretation conditions), even though the voiceless tokens (lacking the preceding word altogether) are much shorter in duration. It is presumably not the case that the vowel recordings from Experiment 1 are simply much more difficult to discriminate because they are from the same speakers, and because the same vowel tokens in the added voiced condition (*ride*) have discrimination accuracy that is 5.6 percentage points higher than the main experimental tokens. Another reason to expect that the preceding word durational difference is not the reason for the observed discrimination difference comes from Experiment 4b, where (despite the same durational differences as in Experiment 4 here) there is no advantage for the two-interpretation condition.

Chapter 5

Importance of Input

This chapter presents a series of experiments that are semi-replications of the first four experiments of this dissertation, with the exception of certain experimental conditions (particularly the voiced environment, words like *ride*) removed. These are Experiments 1b, 1c, 2b, 3b, and 4b, and they consistently (with one informative exception in Experiment 4b) fail to replicate the original findings about the effect of partial contrast on discrimination. These results suggest that the effects of partial contrast on discrimination (discussed in previous chapters) are sensitive to, and in fact reliant on, the listener having a certain quality and/or quantity of exposure to the speakers' voices, a finding that will be examined at the end of this chapter in a general discussion.[1]

5.1 Experiment 1b

5.1.1 Rationale

This experiment tests whether the flap advantage over the voiceless environment (for /aj/) found in Experiment 1 (the initial experiment that demonstrated a partial contrast effect on discrimination) is still found if the voiced environment is not included in the experiment.

[1] This discovery was not the original hypothesis of all of the experiments in this section. Rather, it arose out of a decision after Experiment 1 to leave out the voiced condition to shorten the experiments and focus on the voiceless and flap conditions. This produced a series of unexpected null results that were investigated and explained by re-testing with a voiced environment added or removed (depending on the original experiment). These experiments have been grouped and ordered for a more coherent topic-based (rather than purely chronological) presentation in this dissertation.

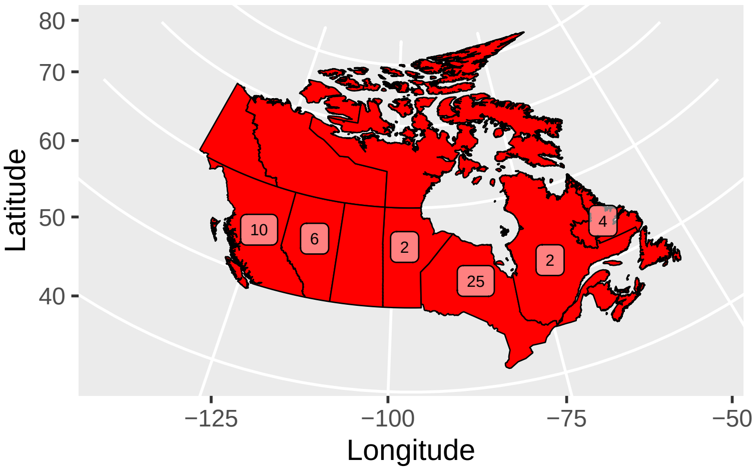

Figure 5.1: Distribution of Participants in Experiment 1b (province of birth)

Table 5.1: Experiment 1b stimuli

Voiceless	Flap (Raised)	Flap (Non-raised)
write	writing	riding
sight	sighting	siding
clout	clouting	clouding
doubt	doubting	Dowding

5.1.2 Method

5.1.2.1 Participants

Forty-nine native speakers of English (born and currently living in Canada, mean age = 32.7, SD = 9, 26 men and 23 women) were recruited to participate in a browser-based online experiment through Prolific (an additional four participants were tested but excluded for reasons of demographic/language criteria). The geographic distribution of participants is provided in Figure 5.1. Most participants indicated little-to-no experience with linguistics. Participants provided informed consent using an online consent form. Participants were paid CAD$3.40 for a session of approximately 15 minutes.

5.1.2.2 Items

The stimuli in this experiment (Table 5.1) were the exact same tokens as in Experiment 1, but with the tokens for the voiced environment removed, leaving only the voiceless and flap environments. As before,

Table 5.2: Mixed effects logistic regression model for Experiment 1b

term	estimate	std.error	statistic	p.value
(Intercept)	0.506	0.051	9.978	0.000
environmentflap	0.093	0.060	1.543	0.123
diphthongaw	-0.462	0.059	-7.858	0.000
environmentflap:diphthongaw	-0.078	0.118	-0.663	0.507

the flap environment did not have any more stimuli than the voiceless environment, but flap environment stimuli were chosen with the intention that raised and non-raised vowels would create different words. This was more successful for /aj/ than /aw/.

5.1.2.3 Procedure

This experiment mirrored the online version of Experiment 1, except that there were 128 trials in Experiment 1b, compared to the original 192.

5.1.3 Results

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 5.2. In contrast to Experiment 1, which had a 9.4 percentage point flap advantage (over the voiceless condition) in the /aj/ diphthong and a 0.3 percentage point advantage in the /aw/ diphthong, in Experiment 1b there was a 2.7 percentage point flap advantage for /aj/ and a 1.2 percentage point flap advantage for /aw/.

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The fixed effects were environment (two levels: voiceless and flap) and diphthong (two levels: aj and aw). The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference levels for environment and diphthong "voiceless" and "aj", respectively. The approach to random effects was to mirror the structure used in Experiment 1, which resulted in by-subjects and by-items random intercepts, as well as by-subjects random slopes for environment.

The one significant result in this model involved diphthong, due to overall higher accuracy for /aj/ (67.3

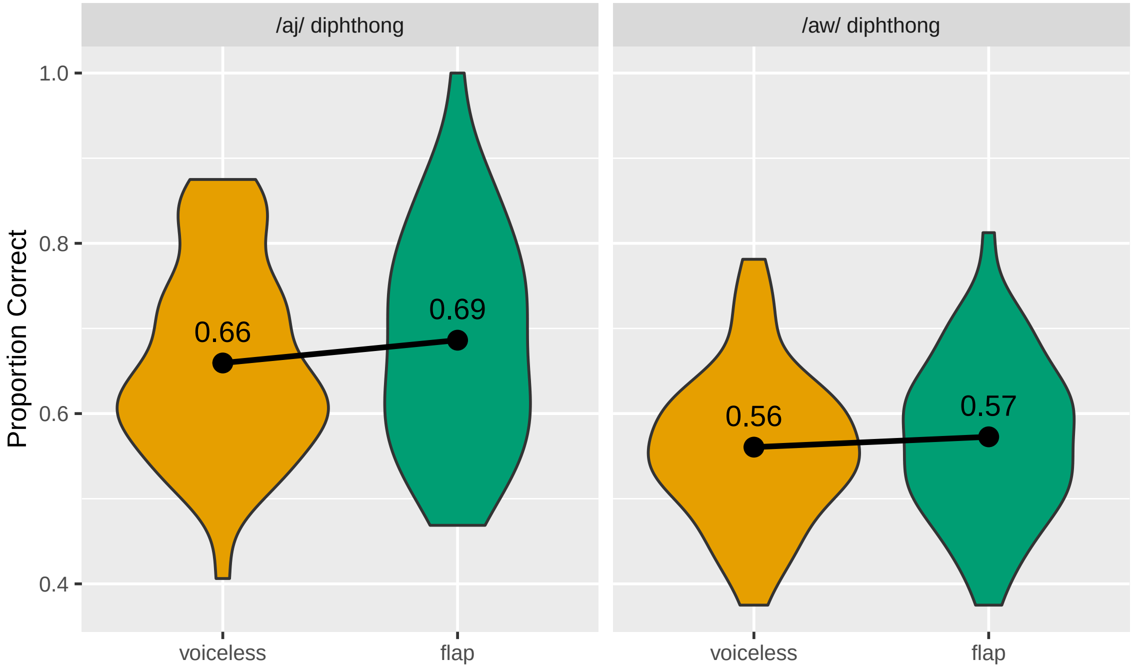

Figure 5.2: Discrimination of diphthong variants in Experiment 1b

percent) than for /aw/ (56.7 percent). In contrast to Experiment 1, where there was a significant interaction between environment and diphthong that indicated a flap advantage for the /aj/ diphthong but not the /aw/ diphthong, in Experiment 1b there is no such interaction. Thus, with the removal of the voiced environment tokens, the primary finding of a flap advantage in Experiment 1 does not appear to replicate.

5.2 Experiment 1c

5.2.1 Rationale

This experiment tests whether the flap advantage (for /aj/) found in Experiment 1 (the initial experiment that demonstrated a partial contrast effect on discrimination) can be further reduced or eliminated by the removal of the /aw/ diphthong tokens.

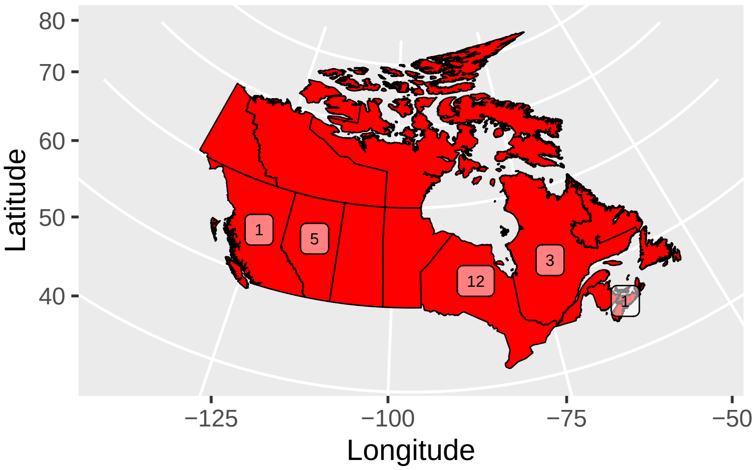

Figure 5.3: Distribution of Participants in Experiment 1c (province of birth)

Table 5.3: Experiment 1b stimuli

Voiceless	Flap (Raised)	Flap (Non-raised)
write	writing	riding
sight	sighting	siding

5.2.2 Method

5.2.2.1 Participants

Twenty-two native speakers of English (born and currently living in Canada, mean age = 30.1, SD = 9.6, 10 men and 12 women) were recruited to participate in a browser-based online experiment through Prolific (an additional two participants were tested but excluded for reasons of demographic/language criteria). The geographic distribution of participants is provided in Figure 5.3. Most participants indicated little-to-no experience with linguistics. Participants provided informed consent using an online consent form. Participants were paid CAD$2.00 for a session of approximately 8 minutes.

5.2.2.2 Items

The stimuli in this experiment (Table 5.3) were the exact same tokens as in Experiment 1, but with the tokens for the voiced environment removed and /aw/ tokens removed, leaving only the voiceless and flap environments and only for /aj/.

Table 5.4: Mixed effects logistic regression model for Experiment 1c

term	estimate	std.error	statistic	p.value
(Intercept)	0.937	0.090	10.378	0.000
environmentflap	0.123	0.124	0.993	0.321

5.2.2.3 Procedure

This experiment mirrored the online version of Experiment 1, except that there were 64 trials in Experiment 1c, compared to 128 in Experiment 1b and 192 in the original Experiment 1.

5.2.3 Results

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 5.2. In contrast to Experiment 1, which had a 9.4 percentage point flap advantage (over the voiceless condition) in the /aj/ diphthong, and Experiment 1b where there was a 2.7 point flap advantage in /aj/, in Experiment 1c there was a 2.1 point flap advantage in /aj/.

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The one fixed effect was environment (two levels: voiceless and flap). The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level for environment was "voiceless". The approach to random effects was to mirror the structure used in Experiment 1, which resulted in by-subjects and by-items random intercepts, as well as by-subjects random slopes for environment. The results of the mixed effects analysis are presented in Figure 5.2.

There was no significant effect of the only predictor variable, environment, in the model, which means that no significant flap advantage was found.

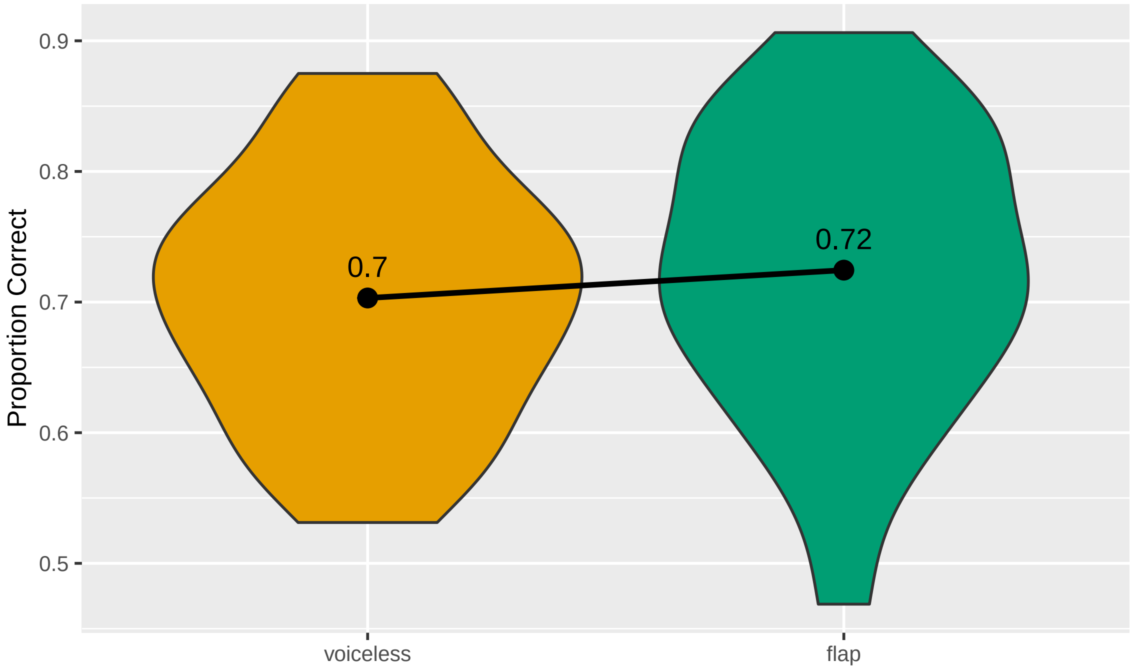

Figure 5.4: Discrimination of diphthong variants in Experiment 1c

5.3 Experiment 2b

5.3.1 Rationale

This experiment tests whether the diphthong-dependent difference between the flap and voiceless environments found in Experiment 2 (the follow-up to Experiment 1 that tested a similar design with non-words) is still found if the voiced environment is not included in the experiment.

5.3.2 Method

5.3.2.1 Participants

Forty-seven native speakers of English (born and currently living in Canada, mean age = 30.6, SD = 10.6, 28 men, 18 women, and 1 other) were recruited to participate in a browser-based online experiment through Prolific (an additional four participants were tested but excluded for reasons of demographic/language criteria). The geographic distribution of participants is provided in Figure 5.5. Most participants indicated little-to-no experience with linguistics. Participants provided informed consent us-

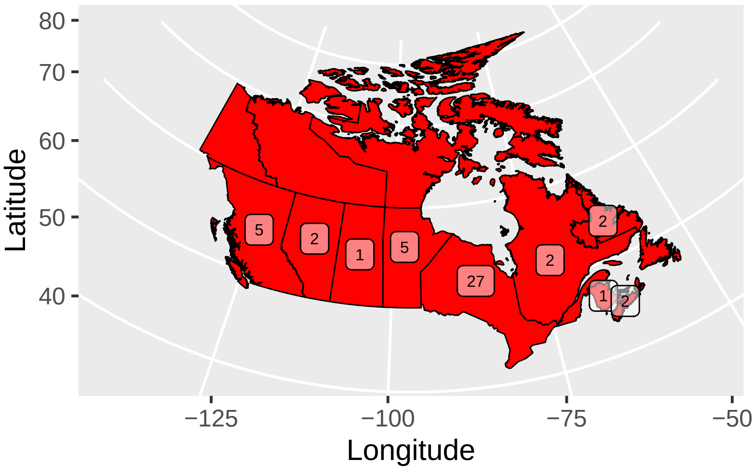

Figure 5.5: Distribution of Participants in Experiment 2b (province of birth)

Table 5.5: Experiment 2b stimuli

Voiceless	Flap
fidight	fidighting
kuvight	kuvighting
stazight	stazighting
fidaut	fidauting
kuvaut	kuvauting
stazaut	stazauting

ing an online consent form. Participants were paid CAD$3.00 for a session of approximately 12 minutes.

5.3.2.2 Items

The stimuli in this experiment (Table 5.5) were the exact same tokens as in Experiment 2, but with the tokens for the voiced environment removed, leaving only the voiceless and flap environments.

5.3.2.3 Procedure

This experiment mirrored Experiment 2, except that there were 96 trials in Experiment 2b, compared to 144 in the original Experiment 2.

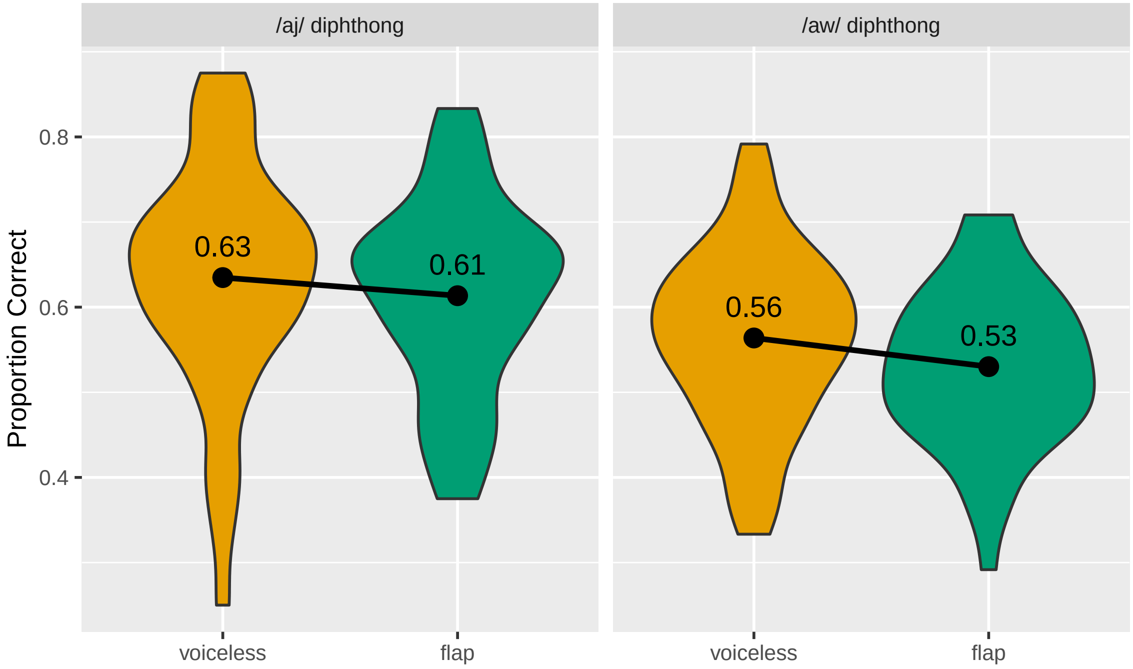

Figure 5.6: Discrimination of diphthong variants in Experiment 2b

5.3.3 Results

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 5.6. In contrast to Experiment 2, which had a 3.4 percentage point flap advantage in /aj/ and a 2.1 point flap disadvantage in /aw/, in Experiment 2b there was a 2.1 point flap disadvantage in /aj/ and a 2.8 point flap disadvantage in /aw/.

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The fixed effects were environment (two levels: voiceless and flap) and diphthong (two levels: aj and aw). The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level for environment was "voiceless" and the reference level for diphthong was "aj". The approach to random effects was to use the same structure as in Experiment 2, which meant by-subjects and by-items random intercepts, as well as by-subjects and by-items random slopes for environment. The items were the rows in Table 5.5, which shared the same onset and the same vowel tokens. The results of the mixed effects analysis are presented in Figure 5.6.

Table 5.6: Mixed effects logistic regression model for Experiment 2b

term	estimate	std.error	statistic	p.value
(Intercept)	0.353	0.060	5.877	0.000
environmentflap	-0.115	0.063	-1.831	0.067
diphthongaw	-0.323	0.061	-5.286	0.000
environmentflap:diphthongaw	-0.046	0.122	-0.373	0.709

The one significant result in this model involved diphthong, due to overall higher accuracy for /aj/ (62.3 percent) than for /aw/ (54.4 percent). In contrast to Experiment 2, where there was a significant interaction between environment and diphthong due to more of a flap advantage in /aj/ than /aw/, in Experiment 2b there is no such interaction. Thus, with the removal of the voiced environment tokens, the primary finding about the diphthong-dependent flap advantage in Experiment 2 did not replicate.

5.4 Experiment 3b

5.4.1 Rationale

This experiment tests whether the diphthong-dependent difference between the flap and voiceless environments found in Experiment 3 (the experiment that set aside the /aw/ diphthong to focus on lexical manipulations with /aj/) is still found if the voiced environment is not included in the experiment.

5.4.2 Method

5.4.2.1 Participants

Forty-six native speakers of English (born and currently living in Canada, mean age = 30.8, SD = 7.4, 27 men and 19 women) were recruited to participate in a browser-based online experiment through Prolific (an additional five participants were tested but excluded for reasons of demographic/language criteria). The geographic distribution of participants is provided in Figure 5.7. Most participants indicated little-to-no experience with linguistics. Participants provided informed consent using an online consent form. Participants were paid CAD$3.00 for a session of approximately 12 minutes.

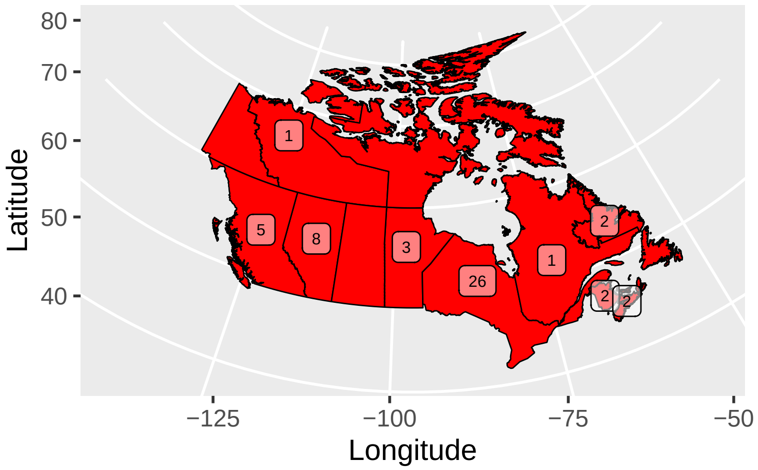

Figure 5.7: Distribution of Participants in Experiment 3b (province of birth)

Table 5.7: Experiment 3b stimuli

Voiceless	Flap
write	writer (rider)
tight	title (tidal)
light	lighter
fight	fighter

5.4.2.2 Items

The stimuli in this experiment (Table 5.7) were the exact same tokens as in Experiment 3, but with the tokens for the voiced environment removed, leaving only the voiceless and flap environments.

5.4.2.3 Procedure

This experiment mirrored Experiment 3, except that Experiment 3b was not split up into two versions to be run on separate participants. For consistency with Experiment 3, the two halves of the stimuli (which were tested on different subjects in Experiment 3 and the same subjects in Experiment 3b) will still be referred to here as sub-experiments. There were 96 trials in Experiment 3b, with 48 in each environment (voiceless and flap). There were also 96 trials in Experiment 3, although there were 32 in each environment (voiceless, flap, and voiced).

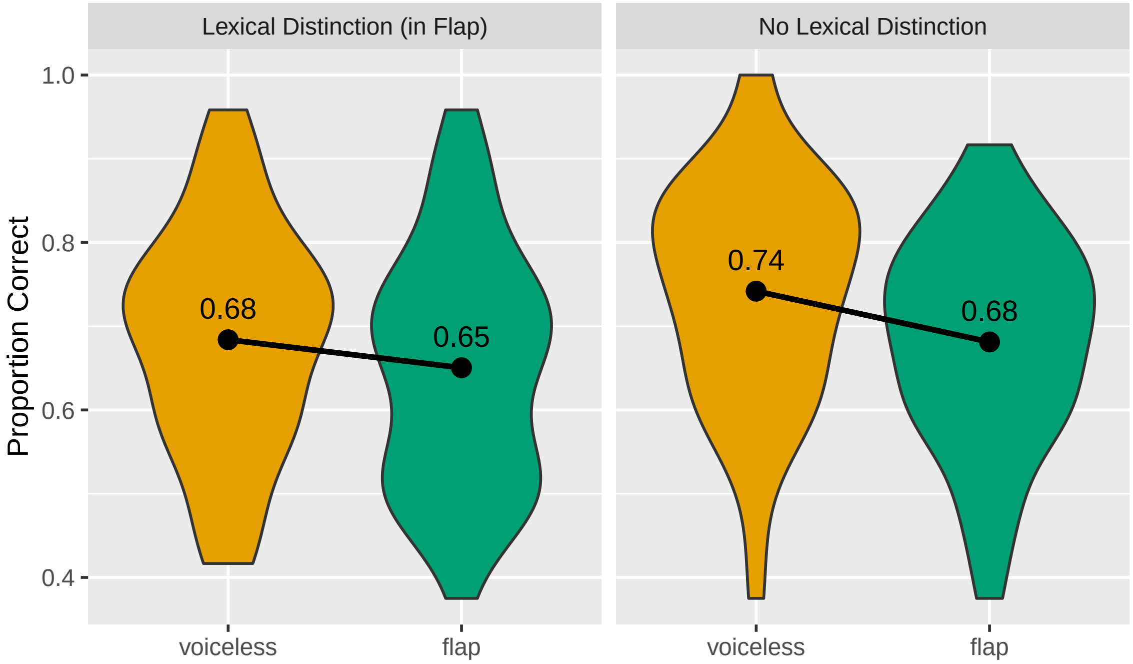

Figure 5.8: Discrimination of diphthong variants in Experiment 3b

5.4.3 Results

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 5.8. In contrast to Experiment 3, which had a 1.5 percentage point flap advantage in the lexical distinction version of the experiment and a 6.9 percentage point flap disadvantage in the no lexical distinction version of the experiment, in Experiment 3b there was a 3.4 point flap disadvantage in the lexical distinction sub-experiment and a 6.1 point flap disadvantage in the no lexical distinction sub-experiment.

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The fixed effects were environment (two levels: voiceless and flap) and sub-experiment (two levels: lexical distinction and no lexical distinction). The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level for environment was "voiceless" and the reference level for sub-experiment was the no lexical distinction sub-experiment. The approach to random effects was to use the same structure as in Experiment 3, which meant by-subjects and by-items random intercepts, with no random slopes. The results of the mixed effects analysis are presented in Figure 5.8.

In contrast to Experiment 3, where the difference between the flap environment and voiceless environ-

Table 5.8: Mixed effects logistic regression model for Experiment 3b ('subexplexdis' refers to the lexical distinction versus no lexical distinction sub-experiments)

term	estimate	std.error	statistic	p.value
(Intercept)	0.757	0.243	3.112	0.002
environmentflap	-0.163	0.094	-1.737	0.082
subexpnolexdis	0.220	0.327	0.674	0.500
environmentflap:subexpnolexdis	-0.153	0.135	-1.135	0.256

ment dependend on the sub-experiment (lexical distinction vs. no lexical distinction) as indicated in a significant interaction, no such interaction was found in Experiment 3b. Thus, with the removal of the voiced environment tokens, the primary finding about the sub-experiment-dependent flap advantage in Experiment 3 did not replicate.

5.5 Experiment 4b

5.5.1 Rationale

This experiment tests whether the advantage in discrimination of the *writing/riding* contrast found for preceding words that allow both interpretations (like *maybe* and *always*) compared to preceding words that prefer one interpretation (like *book* and *train*) in Experiment 4 is still found if the additional voiceless and voiced environment tokens are not included in the experiment.

5.5.2 Method

5.5.2.1 Participants

Twenty-nine native speakers of English (born and currently living in Canada, mean age = 30, SD = 8 . 4, 18 men and 11 women) were recruited to participate in a browser-based online experiment through Prolific (an additional participant was tested but excluded for reasons of demographic/language criteria). The geographic distribution of participants from Prolific is provided in Figure 5.9. Most participants indicated little-to-no experience with linguistics. Participants provided informed consent using an online consent form. Participants were paid CAD$1.50 for a session of approximately 5 minutes.

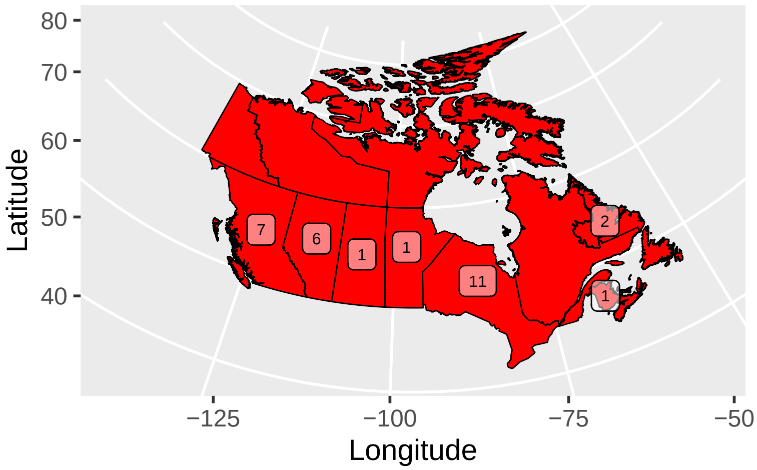

Figure 5.9: Distribution of Participants from Prolific in Experiment 4b (province of birth)

Table 5.9: Experiment 4b stimuli (with COCA co-occurrences) (repeated)

Preceding word	Condition	Preferred	Writing	Riding
Fiction	One-Interpretation	Writing	208	0
Book(s)	One-Interpretation	Writing	245	1
Horseback	One-Interpretation	Riding	0	624
Train(s)	One-Interpretation	Riding	0	13
Still	Two-Interpretation		157	87
Just	Two-Interpretation		172	89
Maybe	Two-Interpretation		14	4
Always	Two-Interpretation		58	19

An additional twenty participants were recruited to participate in this experiment from the friend networks of the two individuals whose voices have been the source of the stimuli for this experiment (and the other experiments in this dissertation).[2] These participants (who took part in the study on a volunteer basis, without monetary compensation) were from Ontario (12), Nova Scotia (5), and British Columbia (2). Due to the difference between these two sources of participants in terms of previous exposure to the speakers' voices, source of participants will be an additional factor in the analysis.

5.5.2.2 Items

The stimuli for Experiment 4b were the exact same as those in Experiment 4, built around the *writing/riding* contrast with the preceding words in Table 4.4 that were intended to encourage one of the interpretations

[2] The twenty participants from the friend group were recruited to pilot test this experiment. The experiment was then put on Prolific to reach the sample size goal of 50, and afterwards it was found that the two groups of participants behaved differently.

(*writing* or *riding*) or to allow both interpretations. The only difference is that the additional voiceless and voiced environment stimuli (*right* and *ride*) from Experiment 1 were not present in Experiment 4b, as they were in Experiment 4.

5.5.2.3 Procedure

This experiment mirrored Experiment 4, except that there were 40 trials instead of 80 due to the lack of additional voiceless and voiced experimental tokens. In addition, the twenty people tested on this experiment that were recruited from the friend networks participated on a version of the experiment that was implemented on the online experiment platform Testable (www.testable.org), rather than through jsPsych. The only difference that this had on the participant experience was that the inter-stimulus interval was set to 750 ms, as opposed to the 1000 ms originally programmed in OpenSesame for the in-lab version of Experiment 1 (which was approximated by a stimulus onset asynchrony, or SOA, of 1500 ms in all of the jsPsych online experiments).

5.5.3 Results

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 5.10. In Experiment 4, there was a 3.4 point discrimination advantage in the two-interpretation condition. In Experiment 4b, there was a 1.9 point disadvantage for the two-interpretation condition among those recruited from Prolific, and a 5.3 point advantage for the two-interpretation condition among those recruited from the friend groups.

The results were analyzed with a mixed effects logistic regression using R (R Core Team, 2017), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0). The fixed effects were interpretations (two levels: one and two) and recruitment (two levels: prolific and friend group). The contrast coding used for this categorical variables was simple coding. The reference level for interpretations was "one" and for recruitment was "prolific". The approach to random effects was to use the same structure as in Experiment 4, which meant by-subjects and by-items random intercepts, with by-subjects random slopes for interpretations. The results of the mixed effects analysis are presented in Figure 5.10.

There was no overall difference between the two-interpretation and one-interpretation conditions, but

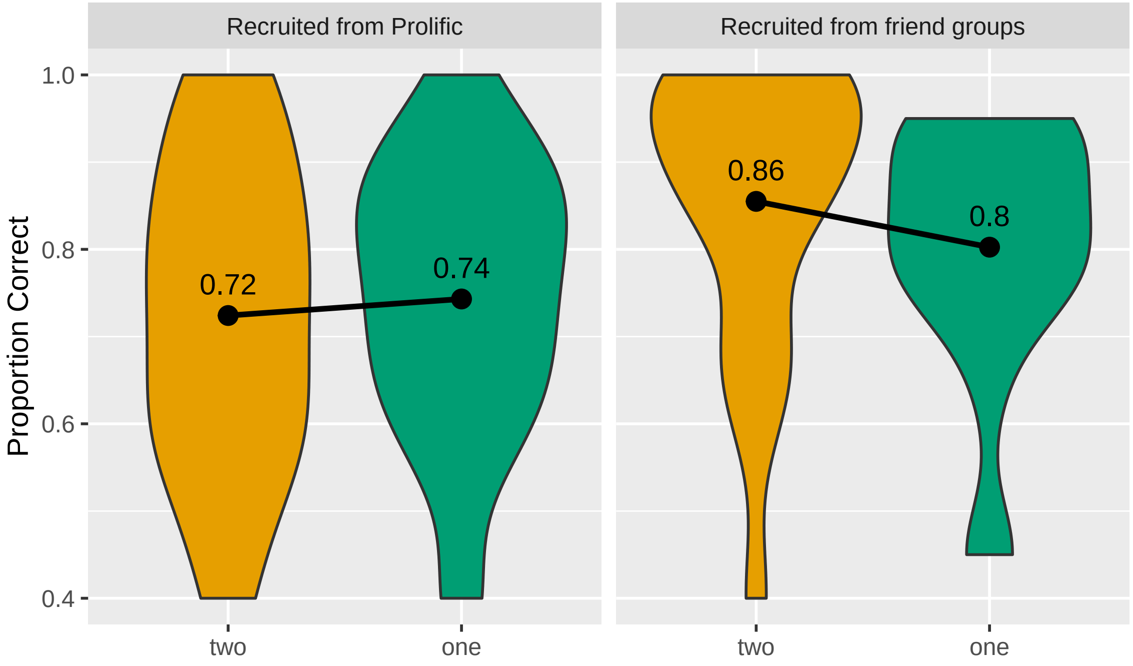

Figure 5.10: Discrimination of diphthong variants in Experiment 4b

Table 5.10: Mixed effects logistic regression model for Experiment 4b

term	estimate	std.error	statistic	p.value
(Intercept)	1.161	0.218	5.313	0.000
interpretationsone	0.012	0.157	0.079	0.937
recruitmentfriendgroup	1.010	0.358	2.823	0.005
interpretationsone:recruitmentfriendgroup	-0.604	0.264	-2.288	0.022

there was a significant interaction between condition and recruitment. This is a result of those recruited from friend groups showing an advantage for the two-interpretation condition, and those recruited from Prolific showing a small disadvantage. According to a post-hoc paired *t*-test, the 5.3 point advantage for the two-interpretation condition among those recruited from friend groups was statistically significant ($t19 = 2.47$, $p = 0.02$). In addition, there was a main effect of recruitment, due to overall better discrimination by those recruited from friend groups (82.9 percent) than those recruited from Prolific (73.4 percent).

5.6 General Discussion

The hypothesis when implementing Experiment 1 was that partial contrast would affect discrimination accuracy, and that this would take the form of better discrimination of raised and non-raised diphthongs in the flap environment (where these variants can and do create different words) than in the voiceless or voiced environments (where these variants cannot create different words, at least in the relevant raising dialects). In addition to the unexpected relevance of diphthong, which was investigated in Experiments 2, 3, and 4, the initial experiment also found unexpectedly high discrimination for the voiced environment (given its status as a non-contrastive environment). It patterned more like the flap environment (which exhibited a discrimination advantage over the voiceless environment, at least for /aj/) than the voiceless environment. However, the voiced environment differs from the other two environments in certain ways, most notably in its substantially longer vowel duration (which can very plausibly explain the higher discrimination), and thus the voiced environment was set aside as being less easily comparable to the other two environments and less relevant for understanding the effect of partial contrast on discrimination. The voiceless and flap environments are more easily comparable. But, as the results in this chapter demonstrate, the voiced environment does appear to be (indirectly) very relevant for the effect of partial contrast on discrimination, because the relationship between the voiceless and flap environments depends not just on factors like diphthong (/aj/ or /aw/) or lexicality (whether or not the raised and non-raised variants create a minimal pair in the given context) but also on the presence or absence of the voiced environment in the experimental context.

In Experiment 1, the relationship between the voiceless and flap environments was dependent on the diphthong, as shown by a significant interaction. There was a sizeable and statistically significant advantage for the flap environment over the voiceless environment for /aj/, but a negligible advantage for the flap environment over the voiceless environment for /aw/. Experiment 1b saw the removal of the voiced environment, and this significant interaction disappeared. Experiment 1c saw the additional removal of the /aw/ diphthong tokens from the experiment, and found still no significant advantage for the flap environment over the voiceless environment for the remaining /aj/ tokens.

Experiment 2 was the version of Experiment 1 with non-words, designed to test the effect of diphthong in the complete absence of lexicality and minimal pairs. Like Experiment 1, it found a significant interaction indicating that the relationship between the voiceless and flap environments was dependent on the diphthong (more of a flap advantage in /aj/ than in /aw/), although post-hoc tests did not find that the difference between the flap and voiceless environments in /aj/ was itself statistically significant. Experiment 2b tested this exact same design but with the removal of the voiced environment from the experiment, and there was no statistically significant interaction indicating that the difference between the voiceless and flap environments depended on the diphthong.

Experiment 3 focused on the /aj/ diphthong and intended to provide a flap condition with minimal pairs and a flap condition without minimal pairs, designed to test the effect of lexicality and minimal pairs. Like the previous two experiments, a significant interaction was found, indicating that the flap condition with minimal pairs had a stronger position (relative to its voiceless comparison condition) than the flap condition without minimal pairs (relative to its voiceless comparison condition). Experiment 3b tested this design without the voiced environment present in the experiment and again found no such significant interaction.

Experiment 4 provided an alternative method of testing the effect of lexicality on discrimination, by testing discrimination of the same flap minimal pair (*writing/riding*) with preceding words that either strongly preferred one interpretation (like *book* and *train*) or allowed both interpretations (like *still* and *just*). It included the voiceless and voiced environment stimuli from Experiment 1, and it found a small but (just barely) statistically significant advantage for the two-interpretation flap condition over the one-interpretation flap condition. Experiment 4b did not include these additional voiceless and voiced environment tokens, and it tested participants from two sources: the same online subject pool that was used for all other online experiments, and the friend groups of the speakers who made the stimuli. The relation between the two-interpretation and one-interpretation condition depended on the group tested (as seen in a significant interaction), with only those recruited from the friend group showing significantly higher discrimination of the two-interpretation condition.

Table 5.11 summarizes these results by comparing the closest equivalent findings across all different experiments. For Experiments 1 and 2 (and their replications), it provides the flap advantage (the difference between the flap and voiceless environments) in /aj/ and in /aw/, and then the interaction, which is the difference between the flap advantage in /aj/ and the flap advantage in /aw/. For Experiment 3 (and its replication), we can see the flap advantage for the flap condition that involves minimal pairs, the flap advantage for the flap condition that does not involve minimal pairs, and the interaction (the difference between these two flap advantages). For Experiment 4 (and its replication), the advantage for the two-interpretation flap condition over the one-interpretation flap condition is indicated.

The absolute flap advantage (the simple difference between discrimination accuracy in the flap environment compared to the voiceless environment) was the original indicator of the effect of partial contrast on discrimination as laid out in the hypothesis for Experiment 1 (and used in the evaluation of the results for Experiment 1).[3] However, the interaction is perhaps a better measure to look at for isolating the effect of the experimental manipulations and understanding the nature of the connection between partial contrast and discrimination. This distinction did not matter for Experiment 1, where the absolute flap advantage for /aj/ was 9.4 points and the relative flap advantage for /aj/ was a very similar 9.1 points (due to the flap advantage for /aw/ being very close to 0), but for later experiments this matters more. Experiment 2 discarded lexicality (by using non-words) to determine the effect of diphthong (i.e., /aj/ vs. /aw/) on the flap advantage. The effect of diphthong that was discovered is not the 3.4 point absolute flap advantage found for /aj/, but rather the difference between the flap advantage for /aj/ and the flap advantage (actually a flap disadvantage) for /aw/, which is 5.5 points. Similarly, in Experiment 3 (which set aside the /aw/ diphthong to test the effect of lexicality in the /aj/ diphthong), the effect of lexicality was the 5.4 point interaction found—having minimal pairs in the flap condition increased performance (relative to the voiceless condition) by 5.4 points.

Table 5.11: Main findings in experiments on partial contrast

Experiment	Finding 1	Finding 2	Interaction
	Flap Adv /aj/	Flap Adv /aw/	
Experiment 1 (!)	**9.4**	**0.3**	**9.1**
Experiment 2 (!)	**3.4**	**-2.1**	**5.5**
Experiment 1b	2.7	1.2	1.5

[3] As mentioned at the beginning of this discussion, the hypothesis also involved comparing the flap environment to the voiced environment, although since Experiment 1 the focus has been on the flap advantage compared to the voiceless environment.

Experiment	Finding 1	Finding 2	Interaction
Experiment 1c	2.1		
Experiment 2b	-2.1	-2.8	0.7
	Flap Adv MP	Flap Adv NoMP	
Experiment 3 (!)	**1.5**	**-3.9**	**5.4**
Experiment 3b	-3.4	-6.1	2.7
	Adv 2-Interp		
Experiment 4 (!)	**3.4**		
Experiment 4b (familiar)	**5.3**		
Experiment 4b (non-fam)	-1.9		

Bolded rows in Table 5.11 indicate experiments where a significant interaction was found, if the design involved analysis of an interaction, or else a significant result for "Finding 1". Exclamations mark experiments that included a voiced condition, which were the original four experiments. As can be seen, the bolded rows line up with the original four experiments that included voiced environment tokens (except for the version of Experiment 4b that was run on people familiar with one or both of the speakers' voices, which also found a significant effect). This reliance on the voiced condition might plausibly be explained by the voiced condition providing listeners with extra information to create models of each speaker's vowel space and phonological patternining, which could provide special benefit for the discrimination of raised and non-raised vowels in a lexically contrastive environment (*writing/riding*, etc.). The discrepancy between the two participant groups in Experiment 4b—only the participants familiar with the speakers' voices exhibited an effect—supports this exposure-based explanation. If this is the case, it would be reminiscent of the Familiar Talker Advantage, where listeners perform better on various perceptual tasks when dealing with familiar voices (Souza et al., 2013; Case et al., 2018b,a). However, this finding is more complicated because familiarity affects the relation between two experimental conditions, rather than just raising overall performance.

There are two possibilities for why it is the voiced environment that produces these effects. One possibility is that the voiced condition simply provided a greater quantity of exposure to raised and non-raised vowel tokens, and that a similar effect could have been found in the absence of the voiced environment if there were simply more trials from the other experimental conditions. Another possibility is that the key is that the voiced condition provides a specific kind of exposure—specifically, the voiced environment substantially increases the listeners' exposure to non-raised vowels in a licit environment, and it uniquely provides exposure to raised and non-raised tokens that are longer in duration than in other environments (due to the longer vowel duration preceding voiced consonants). One reason to think that the latter explanation, that the voiced environment provides a specific quality of exposure (rather than simply additional quantity), is that the voiced environment (when it is included) has consistently been the one with the highest discrimination. It could be that the longer duration vowels in the voiced condition are easier to process and discriminate, and as such they provide special benefit for the listeners in creating models of the speakers' vowel spaces.

Chapter 3 demonstrated an effect of partial contrast on discrimination by means of higher accuracy in the phonological environments where minimal pairs can be created, and Chapter 4 found an effect of partial contrast on discrimination that took the form of higher accuracy in the lexical environments where minimal pairs actually have been created. These results from Chapter 5 provide a novel finding that these partial contrast effects on discrimination are sensitive to the quantity and/or quality of input and linguistic exposure. This is a finding that should be considered by those researching partial contrast in the future; the effects that they hypothesize might exist but not be found if the listeners do not have adequate exposure to their speakers' voices.

Chapter 6

Cross-Dialectal Perception of Canadian Raising

This chapter presents the final experiment of this dissertation, Experiment 5, which investigates the discrimination of Canadian Raising by two groups of American English speakers (from the U.S. North and U.S. West regions) in addition to Canadian English speakers. The main finding is that regions with greater levels of Canadian Raising—which appears to involve not just production but also exposure—respond both more accurately and more quickly, although there is evidence that the effects on accuracy and on reaction time might have different underlying sources (a diphthong difference in accuracy suggests that partial contrast plays a greater role in the accuracy results than the reaction time results). In addition, the findings of a dialect survey suggest that (1) raising might be more common in the United States, particularly in the U.S. West region, than is indicated in many sources, and (2) [ʌj]-lexicalization is more common among raisers in the U.S. North than in Canada or the U.S. West.

6.1 Experiment 5

6.1.1 Rationale

This experiment tests speakers of Canadian English and speakers of American English on their discrimination of raised and non-raised diphthongs ([ʌj]~[aj] and [ʌw]~[aw]) in three environments. Two of these environments (before a voiceless /t/ and before a voiced /d/) have been investigated in previous experiments, but this experiment also tests discrimination of these diphthongs in a non-lexical isolation environment (without an onset or a coda). This experiment is designed to test two hypotheses.

The first hypothesis is that those with Canadian Raising in their speech for a particular diphthong will have a harder time discriminating raised and non-raised variants of that diphthong (compared to those who lack Canadian Raising for that diphthong). For example, this predicts that those who raise both /aj/ and /aw/ would perform worse than those who raise neither /aj/ nor /aw/. This hypothesis is based on the idea that raised diphthongs are more marked or noticeable for non-raisers than non-raised diphthongs are for raisers, which comes from an apparent asymmetry between these features in stereotypes of the other dialect. Canadian Raising (especially of /aw/) is a stereotypical (and frequently remarked on) identifier of Canadian English often exaggerated as *oot and aboot* (Chambers, 1973; Boberg, 2008). In the other direction, however, non-raising (of either /aj/ or /aw/) does not appear to feature as prominently in the aspects of American English that are remarked on or stereotyped, aside from experiences almost 80 years ago reported in Joos (1942) ("if I use a low diphthong before a fortis consonant [...] the Canadian listener immediately accuses me of drawling", p. 142). If raised diphthongs are in fact more marked or noticeable for non-raisers than non-raised diphthongs are for raisers, it could be because raisers still produce non-raised variants in certain environments, because non-raised diphthongs are the underlying or canonical form of the vowel, or because raisers have more exposure to non-raising dialects in the media. This first hypothesis will be tested on both accuracy and response time as outcome variables. Response time was not included in previous analyses because the length of the stimuli had not been equated across environments (stimuli in the flap environment had an *-ing* suffix, making them longer than stimuli in other environments), but the between-subjects nature of this hypothesis allows the use of response time because stimulus duration does not differ between groups.

The second hypothesis to be tested in this experiment is that speakers of Canadian English will have an easier time discriminating between diphthong variants when they are presented in isolation, compared to being presented in the allophonic (voiceless) context, when the vowel duration is equivalent. This hypothesis is based on findings in past experiments that discrimination of raised and non-raised diphthongs can vary between the contrastive (flap) environment and the allophonic (voiceless) environment; it is thus possible that the isolation context, which is not contrastive but is also not allophonic, might also present an advantage over the allophonic environment.

6.1.2 Method

6.1.2.1 Participants

One hundred and forty eight native speakers of English were recruited to participate in a browser-based online experiment through Prolific (an additional two participants were tested but excluded from analysis for reasons of language background or country status). One-third of these participants (49 people, mean age = 32.5, SD = 11.7, 24 men and 25 women) were born, and are currently living, in Canada. Another one-third of participants (50 people, mean age = 33.8, SD = 10.9, 27 men, 22 women, 1 other) were born, and are currently living, in the region of the United States where /aj/ raising (but not /aw/ raising) is generally expected, which is an area in the Northern United States centred on the Great Lakes but extending as far west as the Dakotas and as far east as parts of New England (Vance, 1987; Allen, 1989; Dailey-O'Cain, 1997; Niedzielski, 1999; Labov et al., 2006). The particular area of eligibility included 11 states (North Dakota, South Dakota, Minnesota, Wisconsin, Michigan, Pennsylvania, New York, Maryland, Massachusetts, New Hampshire, and Rhode Island), and it was based on the the American portion of the "Canadian Raising of /ay/" isogloss (Map 14.10, p. 206) in the *Atlas of North American English* (Labov et al., 2006). This isogloss extends further south to include significant parts of Illinois, Indiana, and Ohio, but these were excluded because it was not possible to target areas within a state. Finally, one-third of participants (49 people, mean age = 31.1, SD = 11.3, 32 men, 14 women, 3 other) were born, and are currently living, in the Western United States. This area is defined by Labov et al. (2006) as sharing many features with Canadian English, but not Canadian Raising. Thus, based on the *Atlas of North American English*, this group is expected to differ from the U.S. North group on one diphthong, and the Canadian group on both diphthongs. However, more recent findings suggest that raising of /aj/ might be more common in the Western United States than is indicated in the *Atlas* [Sadlier-Brown (2012); Swan (2016); Tyler Kendall, personal communication], and so another method of grouping participants (by production) is discussed below. The area of eligibility for the third group included 11 different states (California, Nevada, Arizona, New Mexico, Colorado, Utah, Wyoming, Idaho, Oregon, Washington, and Montana).[1] Figure 6.1 shows the three regions used in recruitment, as well as the distribution of participants according to their province or state of birth. Most participants indicated little-to-no experience with linguistics. Participants provided informed consent using an online consent form. Participants were paid CAD$3.00 for a session

[1] The U.S. South region was avoided, even though raising is generally not expected there, because this region has monopthongization (or glide-deletion) that results in /aj/ beiong realized as closer to [a] (Allbritten, 2011). This could affect discrimination of Canadian Raising diphthongs.

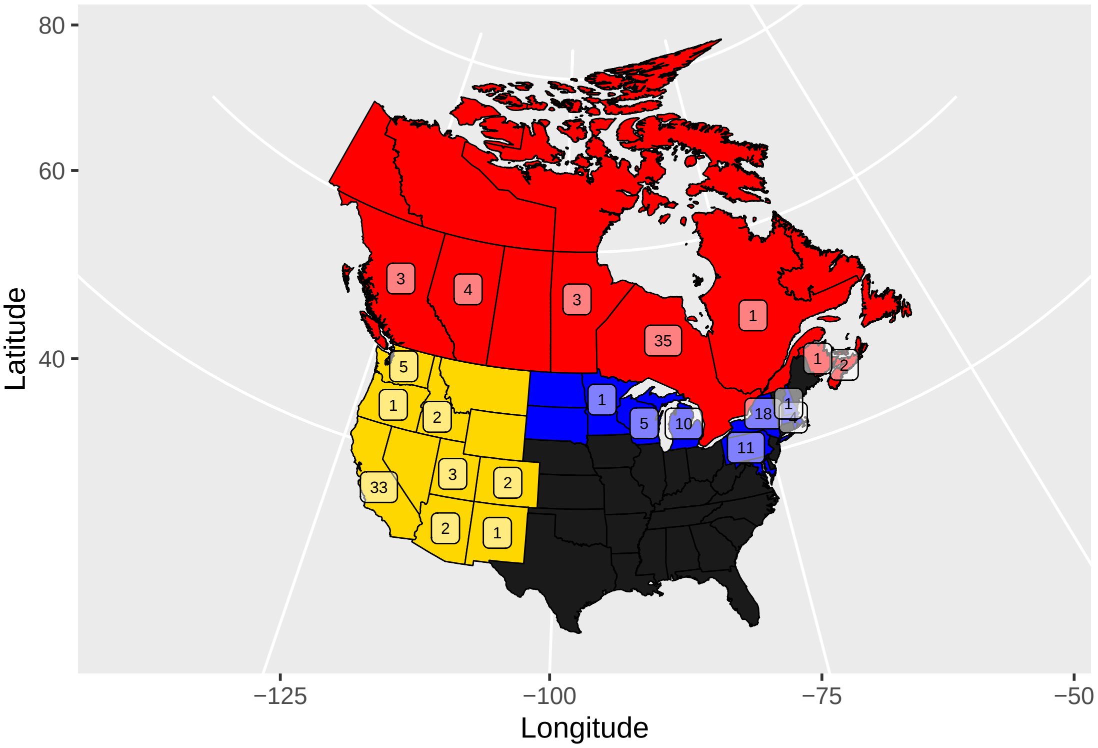

Figure 6.1: Distribution of Participants in Experiment 5 (province/state of birth)

of approximately 12 minutes.

Due to the unclear state of the prevalence of /aj/ raising in the Western United States, this experiment included a short survey at the beginning that asked participants whether (in their most natural pronunciation) the words *rider* and *writer* sound the same or different. Dialect surveys have a history of use in sociolinguistics, for example in the Dialect Topography of Canada project (Chambers, 1994; Chambers and Heisler, 1999; Pi, 2000; Chambers, 2006b). Answering that these words sound different suggests the presence of /aj/ raising,[2] while answering that these words sound the same suggests an absence of /aj/ raising. The counts of the responses by region are shown in Figure 6.1. These counts correspond to an /aj/ raising rate of 84 percent for Canadians, which lines up exactly with an /aj/ raising prevalence of 84 percent found by Boberg (2008) in an acoustic analysis of a geographically-diverse sample of Canadians. The /aj/ raising rate was 80 percent in the U.S. North and 59 percent in the U.S. West. Thus, particularly in the United States, (self-reported) raising does not line up with region in the way that was initially expected based on Labov et al. (2006). This raises the possibility of analyzing the results by region or by production.

[2] It is possible that those without raising would indicate a difference between *rider* and *writer* due to durational differences in the vowel. However, Braver (2013) finds only a small difference of 9 ms between pre-flap vowels depending on whether the flap is underlyingly a /t/ or a /d/, and Braver (2014) finds this difference to be imperceptible.

Table 6.1: Self-reported /aj/ raising among participants in Experiment 5 by region

Response	Canada	U.S. North	U.S. West
Raises /aj/	41	40	29
Does not raise /aj/	8	10	20

Table 6.2: Experiment 5 stimuli

Isolation	Voiceless	Voiced
(non-lexical)	height	hide
(non-lexical)	out	how'd

The results section will include both analyses, and will show that the analysis by region better accounts for the data.

The survey at the beginning included a question for /aj/ raising but not for /aw/ raising, and so some assumptions had to be made about /aw/ raising when grouping participants by production. For Canadians, it was assumed that their /aw/ raising matched their /aj/ raising, meaning that if they reported /aj/ raising then they were treated as raising both, and if they reported no /aj/ raising they were treated as raising neither. For Americans, it was assumed that they did not raise /aw/, and so those who reported /aj/ raising were coded as raising only /aj/, and those who reported no /aj/ raising were coded as raising neither. This categorization resulted in three groups: those assumed to raise both /aj/ and /aw/ (42 Canadians), those assumed to raise only /aj/ (73 Americans), and those assumed to raise neither vowel (33 Americans and Canadians).

In addition to the question about *rider* and *writer* that was asked to determine whether participants exhibit Canadian Raising, a second question asked participants who do exhibit such a difference whether *spider* rhymes with *rider* (as is traditionally expected) or with *writer*. The latter response indicates an [ʌj] in *spider*, which cannot be explained by Canadian Raising (that flap is not an underlying /t/) and is in line with the finding that [ʌj] is lexicalizing in some dialects (being found in contexts not predicted by the Canadian Raising allophonic alternation) and possibly emerging as its own phoneme (Fruehwald, 2008; Vance, 1987). Responses to this question, including differences between regions, will be included in the results section.

Table 6.3: Experiment 5 stimulus vowel durations (ms)

Gender	Voiceless	Voiced
Male	196.1	387.0
Female	224.3	374.3
Average	210.2	380.6

6.1.2.2 Items

The stimuli in this experiment are the 4 words (as well as the vowels in isolation) with an /aj/ or /aw/ diphthing in Table 6.2. The addition of the isolation environment to this experiment introduced a restriction on the other environments to have an /h/ onset or no onset at all, because an onset with a regular consonant would introduce formant transitions on the vowel that would remain and possibly affect perception of the vowel even when the diphthong is played in isolation. For the voiceless and voiced environments, stimulus creation in Experiment 5 was identical to Experiment 1 (using the same two speakers and the same process of splicing, including duration manipulation), except that a greater number of tokens for each word were created in Experiment 5 due to a smaller number of words. In Experiment 1, two raised vowel tokens were taken from two recordings of *sight* and two non-raised vowel tokens were taken from two recordings of *side* (these were then spliced into separate recordings of both *sight* and *side* to create two raised and two non-raised versions of each). In Experiment 5, five raised vowel tokens were taken from five recordings of *height* and five non-raised vowel tokens were taken from five recordings of *hide*. As for the isolation environment, which was not included in the stimuli for any other experiment, the same duration was used as the voiceless environment. The vowel durations in the different environments are in Table 6.3. The raised and non-raised vowel tokens in isolation were accoustically the same as the raised and non-raised vowel tokens in the voiceless environment (*height* and *out*), except that a 25 ms fade in and a 25 ms fade out were applied to the vowel in isolation to make it sound less abrupt.

6.1.2.3 Procedure

This experiment was an AXB discrimination task mirroring the online version of Experiment 1. There were 120 trials, with 40 from each environment (isolation, voiceless, and flap) interspersed.

6.1.3 Results

This results section will cover two outcome variables, accuracy and response time (measured from the appearance of the response screen), and it will include the analysis by region (dividing participants into three groups: Canada, U.S. North, and U.S. West) and by production (dividing participants into three groups: raisers of both /aj/ and /aw/, raisers of only /aj/, and non-raisers). These results were analyzed with mixed effects regression models (logistic regression for accuracy and linear regression for response time) using *R* (R Core Team, 2017), *lme4* (Bates et al., 2015), and *lmerTest* (Kuznetsova et al., 2017). The response variable was binary correct (1) or incorrect (0) for accuracy, while it was continuous for response time. The fixed effects were environment (three levels: isolation, voiceless, and flap), diphthong (two levels: aj and aw), and either participant region or participant raising. The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level was "voiceless" for environment, "aj" for diphthong, "Canada" for participant region, and "both" (i.e., raisers of both /aj/ and /aw/) for participant raising. The approach to random effects was to use the maximal structure justified by the experimental design that does not result in a failure to converge or a singular fit (Barr et al., 2013), but to keep the random effects structure the same between the participant region and participant raising analyses to avoid introducing unnecessary differences for model comparison. The result was by-subjects random intercepts with no random slopes for the accuracy data, and by-subjects random intercepts (with random slopes for diphthong) for the response time data. No random effect for item was included because each diphthong/environment combination only had one item. The goal of performing a model comparison also necessitated the use of maximum likelihood (ML) rather than restricted maximum likelihood (REML) in the regression analysis.

Determining whether participant region or participant raising better accounts for the results can be done using model comparison with the Akaike information criterion (AIC) (Akaike, 1974, 1998; Aho et al., 2014; Kingdom and Prins, 2016). A lower AIC indicates a better model fit, and it is judged according to the AIC difference (ΔAIC) between the two models. Rough guidelines proposed in the literature classify a ΔAIC between models of 0–2 as being weak evidence in favour of the lower AIC model, 4–7 as being considerable evidence, and above 10 as being very strong evidence (Raftery, 1995; Burnham and Anderson, 2004). As seen in Table 6.4, there was very strong evidence in favour of the model based on participant region over the model based on participant raising, both for the accuracy data (ΔAIC = 10.3) and the response time data (ΔAIC = 38.7). Thus, participant region accounts for the results of this experiment, both accuracy and response time, better than participant raising.

Table 6.4: Akaike information criterion (AIC) values for two models for each outcome variable

Outcome Variable	Region Model	Raising Model	ΔAIC
Accuracy	22508	22518	10.3
Response Time	292622	292660	38.7

6.1.3.1 Analysis by Participant Region

6.1.3.1.1 Accuracy

The proportion of trials in which the participants correctly discriminated between diphthongs is shown in Figure 6.2, which shows the performance of each of the three regional groups (Canada, U.S. North, and U.S. West) in the three environments (isolation, voiceless, and voiced) for each of the two diphthongs (/aj/ and /aw/). The violin plots showing the overall distribution of data, which were used in the presentation of the results of the previous experiments, were not included here to avoid visual clutter given the greater complexity of the results in this experiment. The blue dashed line provides the average across the isolation, voiceless, and voiced environments for each diphthong, to allow for easier comparison of larger patterns. Visually, overall performance for /aw/ is similar for all three groups, while the U.S. West group patterns lower than the Canada and U.S. North groups for overall /aj/ performance. Due to the larger number of plots and models being presented in this experiment, the model outputs are available in the Appendix. The results of the mixed effects analysis for accuracy by participant region are presented in the Appendix in Table D.1.

To summarize the significant findings ($p < 0.05$) in the model, in order, accuracy was higher in the voiced environment (69.9 percent) than the voiceless environment (62.3 percent) and higher for the /aj/ diphthong (68.8 percent) than the /aw/ diphthong (61.4 percent), which are two results that mirror previous experiments. There was also a significant interaction between the voiced environment and diphthong, due to a larger voiced advantage over the voiceless environment in the /aw/ diphthong (12.5 percentage points) than in the /aj/ diphthong (2.7 percentage points), which does not have a precedent in previous experiments. There were also significant interactions between the isolation environment and both the U.S.

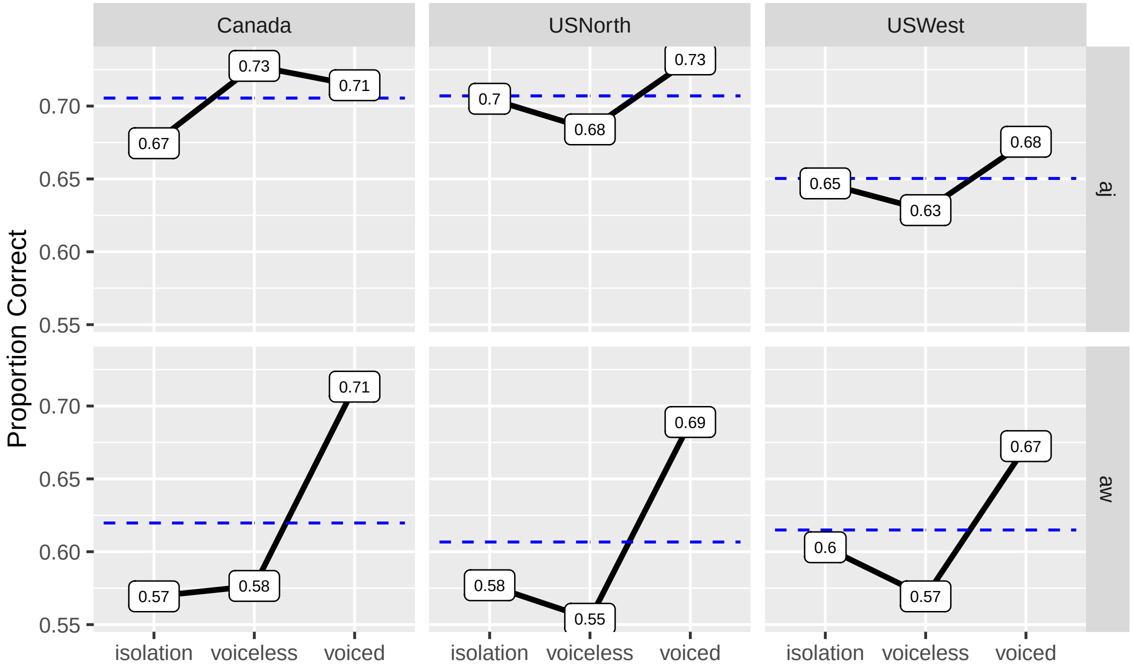

Figure 6.2: Discrimination of diphthong variants in Experiment 5 by participant region (preferred model)

North and U.S. West groups, due to the Canadian group having higher accuracy in the voiceless environment than the isolation environment (by 3 percentage points) but the U.S. North and U.S. West groups having *worse* performance in the voiceless environment than the isolation environments (by 2.2 and 2.6 percentage points, respectively). According to a post-hoc paired *t*-test, the 3 percentage point advantage for the voiceless context over the isolation context exhibited by the Canadians was significant (t_{48} = 2.10, p = 0.04). There was also an interaction between diphthong and the U.S. West region, due to the U.S. West group performing only slightly worse than the Canadian group in the /aw/ diphthong (0.5 percentage points, non-significant in a post-hoc independent t-test, $t_{93.91}$ = 0.25, p = 0.80) but moderately worse on the /aw/ diphthong (4.5 percentage points, significant in a post-hoc independent t-test, $t_{94.12}$ = 2.63, p < 0.01). There was no interaction to indicate the U.S. North group differing from the Canadian group in the same way. Finally, there was a three-way interaction between the voiced environment, diphthong, and the U.S. West region, due to the Canadian group exhibiting a 1.4 percentage point advantage of the voiceless environment over the voiced environment in /aj/ compared to the U.S. West group's 4.7 percentage point disadvantage for the voiceless environment compared to the voiced environment in /aj/ (both groups exhibited a 10–14 point disadvantage for voiceless compared to voiced in the /aw/ diphthong). The Canadian group's 1.4 percentage point advantage for the voiceless environment over the voiced environment in /aj/ was not statistically significant in a post-hoc paired *t*-test (t_{48} = 0.66, p = 0.51).

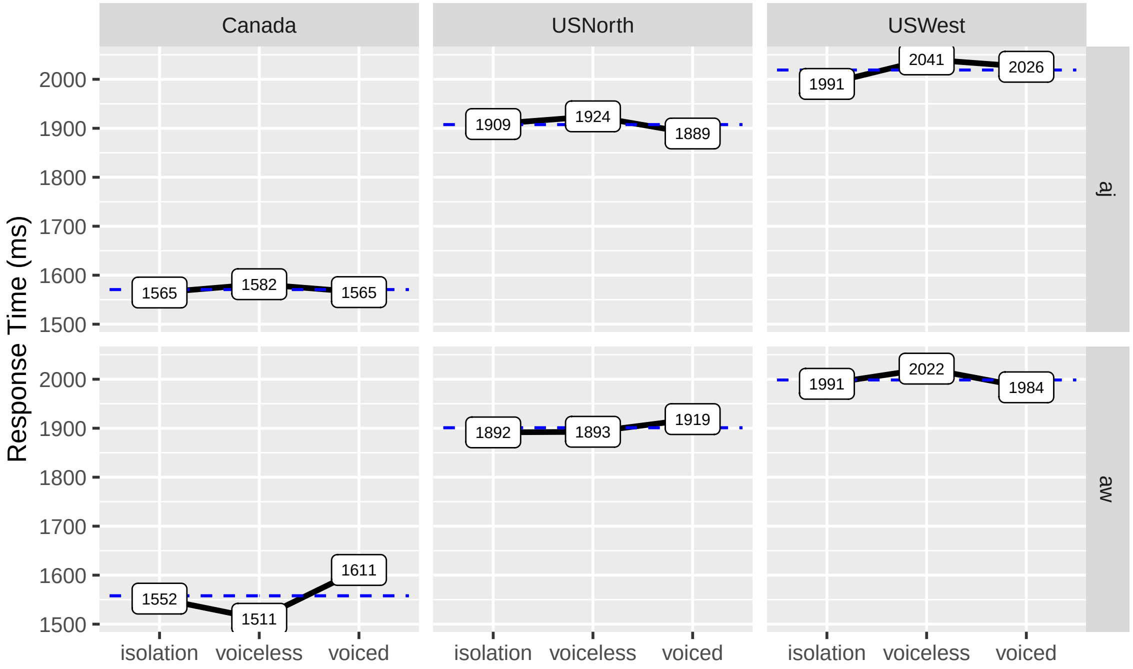

Figure 6.3: Response times for discrimination of diphthong variants in Experiment 5 by participant region (preferred model)

6.1.3.1.2 Response Time

The mean response times (response latencies) according to participant region are shown in Figure 6.3. Visually, the Canadian group performs much faster than both American groups across both diphthongs, while the U.S. North group performs somewhat faster than the U.S. West group. The results of the mixed effects analysis for response time by participant region are presented in the Appendix in Table D.2.

There were only two significant findings ($p < 0.05$) in the model, which indicated that the Canadian group had overall faster responses (1564 ms) than the U.S. North group (1904 ms) and the U.S. West group (2009 ms). According to a post-hoc independent t-test, the U.S. North group was in turn significantly faster than the U.S. West group ($t_{91.781} = 2.78$, $p < 0.01$).

6.1.3.2 Analysis by Participant Raising

6.1.3.2.1 Accuracy

Response accuracy according to participant raising is shown in Figure 6.4. Visually, there is less of a clear pattern compared to the analysis by participant region. The results of the mixed effects analysis for

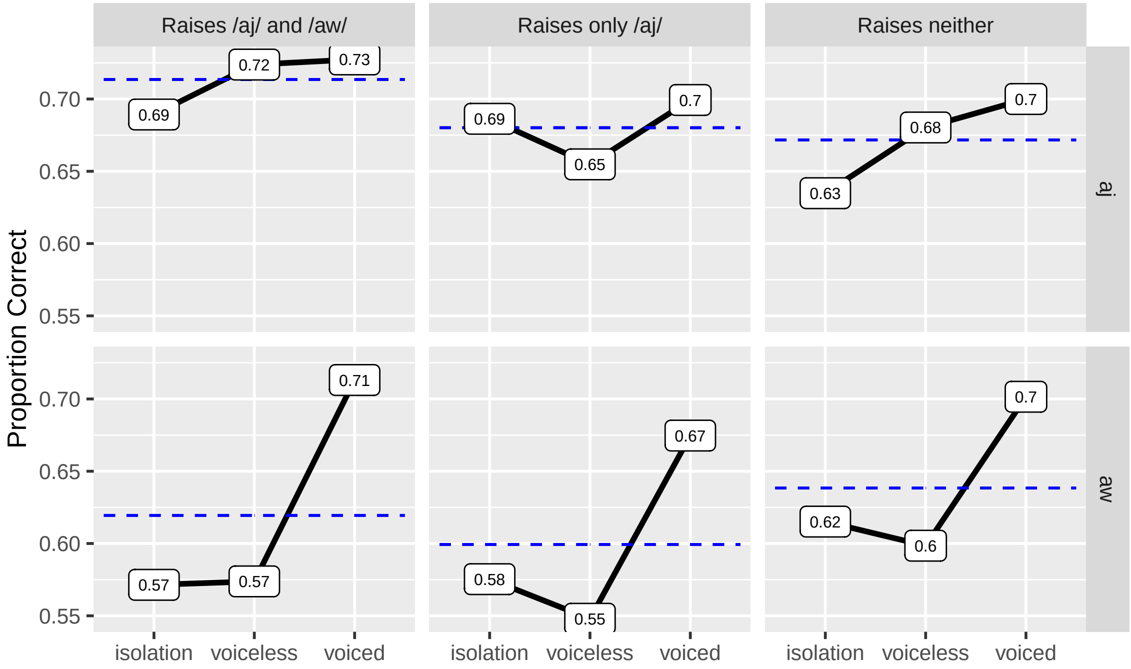

Figure 6.4: Discrimination of diphthong variants in Experiment 5 by participant raising (dispreferred model)

response time by participant region are presented in the Appendix in Table D.3.

Because this model has two of the three same fixed effects as the analysis by participant region, many of the findings will be the same or very similar. This includes the significant results that accuracy was higher in the voiced environment than the voiceless environment, and higher for the /aj/ diphthong than the /aw/ diphthong. It also includes the fact that there was a larger voiced advantage over the voiceless environment in the /aw/ diphthong than in the /aj/ diphthong. There were two findings specific to this analysis (i.e., involving participant raising). First, there was a significant interaction between the isolation environment and the group that raises only /aj/, due to the group that raises both diphthongs having a 1.9 percentage point disadvantage for the isolation environment compared to the voiceless environment (non-significant in a post-hoc independent t-test, t_{41} = 1.17, p = 0.25), while the /aj/-only group exhibited a 3.0 percentage point advantage for the isolation environment over the voiceless environment. Second, there was a significant interaction between diphthong and the group that raises neither, due to the group that raises neither having a 1.9 percentage point advantage over the group that raises both for /aj/, but a 4.1 percentage point disadvantage for /aw/.

6.1.3.2.2 Response Time

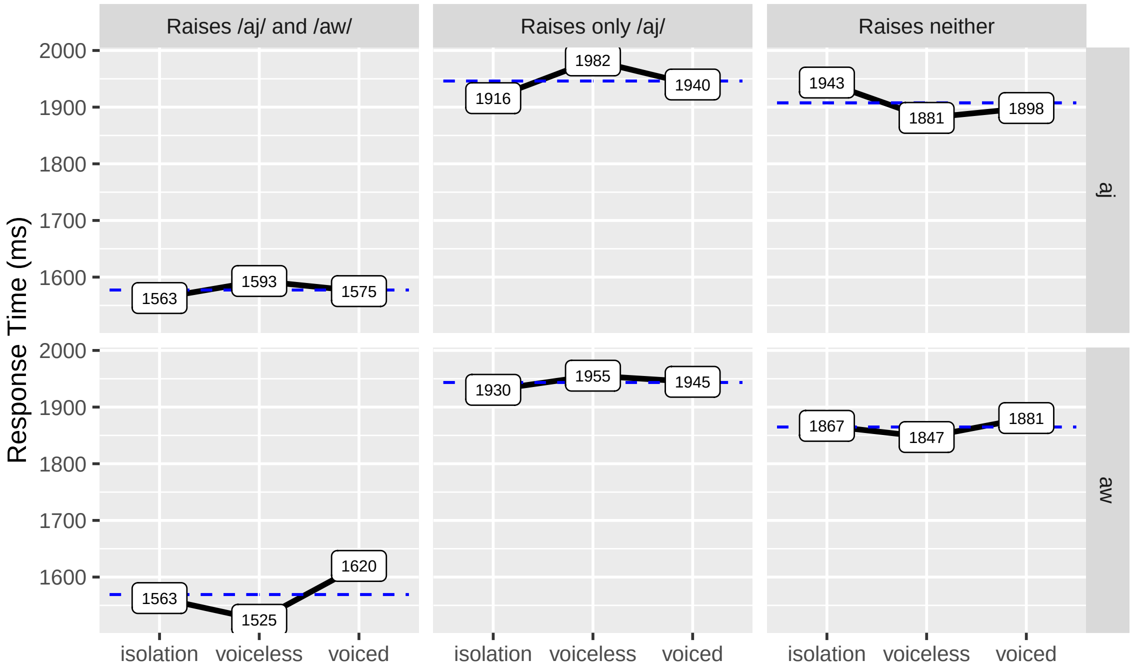

Figure 6.5: Response times for discrimination of diphthong variants in Experiment 5 by participant raising (dispreferred model)

The mean response times according to participant raising are shown in Figure 6.5. Visually, the group that raises both diphthongs performs at a higher accuracy than the other two groups on both diphthongs. The results of the mixed effects analysis for response time by participant region are presented in the Appendix in Table D.4.

There were only two significant findings ($p < 0.05$) in the model, which indicated that the group that raises both diphthongs had overall faster responses (1573 ms) than the group that raises only /aj/ (1945 ms) and the group that raises neither (1886 ms). According to a post-hoc independent t-test, there was no significant difference between the latter two groups ($t_{47.23} = 1.12$, $p = 0.26$).

6.1.3.3 Pronunciation of “spider”

In addition to the behavioural data, participants who self-reported that *rider* and *writer* sound different (which suggests Canadian Raising of /aj/ in their speech) also reported whether *spider* rhymes with *rider* (indicating that they pronounce it with [aj]) or with *writer* (indicating that they pronounce it with [ʌj]). The pronunciation with [ʌj] was most common in the U.S. North (34 percent of participants with raising) and least common in the U.S. West (3 percent of participants with raising), with Canada in the middle

Table 6.5: Self-reported pronunciation of 'spider' by region (among those who self-report having Canadian Raising)

Response	Canada	U.S. North	U.S. West
Spider with [aj]	37	24	27
Spider with [ʌj]	4	13	1
Unsure/blank		1	1

(13 percent of participants with raising). According to a χ^2 Test for Independence (excluding the two unsure/blank responses), there was a significant effect of region on response ($\chi^2{}_2$ = 13.739, p = 0.001).

6.1.4 Discussion

This experiment was designed to test two hypotheses. The first hypothesis was that those with Canadian Raising in their speech for a particular diphthong will have a harder time discriminating raised and non-raised variants of that diphthong, compared to those who lack Canadian Raising in their speech for that diphthong. The second hypothesis was that speakers of Canadian English will have an easier time discriminating between diphthong variants when they are presented in isolation, compared to being presented in the allophonic (voiceless) context.

6.1.4.1 Group Differences in Perception

The first hypothesis would predict a pattern where those who raise both /aj/ and /aw/ have the lowest accuracy (and slowest response times), while those who raise just /aj/ would have elevated performance for /aw/, and those who raise neither would have the highest performance overall. This pattern was not found for either response time or accuracy, and in fact the results are more in the opposite direction, particularly for response time where the group that raises both /aj/ and /aw/ has the fastest performance rather than the slowest. However, a model comparison found strong evidence that the model based on participant region accounted for the results (both accuracy and response time) better than the model based on participant raising, which suggests that region can better explain the discriminability of Canadian Raising diphthong variants than production alone. This raises the question of whether self-reported production results are reliable, but there is some reason to believe that they are in this case, because the self-reported raising prevalence in the sample of Canadians (84 percent identified raising /aj/) matched the acoustic analysis of Boberg (2008), which found that 84 percent of Canadians (in a geographically diverse sample) raise

/aj/. Thus, it appears that participants understood the dialect question and were adequately aware of their pronunciation to answer it accurately.

The analysis by participant region, the better supported model, found for accuracy that all three groups performed at approximately the same level for /aw/, while the Canadians and U.S. North groups had higher accuracy on /aj/ than the U.S. West group (by approximately five percentage points). Looking at response time, Canadians were 340 ms faster (across diphthongs) than the U.S. North group, who in turn were 105 ms faster than the U.S. West group. There are three main points to notice in these results. First, the Canadian group performed overall the best and the U.S. West group the worst, which is the opposite of what was predicted. Second, the U.S. North group patterned more closely with the Canadians for accuracy but more closely with the U.S. West for response time. And third, there was a major group difference by diphthong for accuracy but not for response time.

6.1.4.1.1 Accuracy

The diphthong pattern in the accuracy results suggests a role of partial contrast in those findings, but not in the same sense as the partial contrast effect that was investigated in the previous experiments of this dissertation. That was a partial contrast effect in the sense of a context-specific discrimination difference (comparable to Celata, 2008), where discrimination of a partially contrastive sound pair is better in the contrastive environment than the non-contrastive environment. There is also some evidence that partially contrastive sound pairs are, independent of any effect of context/environment, harder to discriminate than fully contrastive sound pairs and easier to discriminate than allophonic (i.e., non-contrastive) sound pairs (Hume and Johnson, 2003; Stevenson and Zamuner, 2017). This could explain the accuracy findings in this experiment, given evidence in previous experiments in this dissertation that only [ʌj]~[aj] has the status of partially contrastive, while [ʌw]~[aw] does not. It could be that all groups perform at similarly low levels when discriminating [ʌw]~[aw] because this sound pair does not have the status of being partially contrastive to any of the three groups, while the Canadian and U.S. North groups have a particular boost in performance for discriminating [ʌj]~[aj] because this sound pair is partially contrastive for them to a greater extent than it is for the U.S. West group. If this explanation is correct, then the large discrimination advantage for /aj/ over /aw/ that has been found in all of the previous experiments is not just a result of there being a larger phonetic difference between [ʌj]~[aj] than [ʌw]~[aw] (as was found in the phonetic analysis chapter)—it is also a result of a general discrimination boost for [ʌj]~[aj] (separate from the context-specific discrimination boost in the flap environment) as a result of that sound pair being partially contrastive for the Canadian participants tested in all of those experiments.

Using partial contrast to explain the differences between Canada and the U.S. North on one hand and the U.S. West on the other hand raises the question of why the U.S. West would differ from the other two regions in terms of partial contrast. The first answer is the lower prevalence of raising in the U.S. West, because if [ʌj] does not exist in an individual's native phonology then it is unlikely that [ʌj]~[aj] has the status of partially contrastive in their phonology. However, raisers were still a majority in the U.S. West group (just a smaller majority: 59 percent, compared to 80 percent in the U.S. North and 84 percent in Canada), and if the presence or absence of raising in the speech of the participants was the only relevant factor then the alternative model based on participant raising (which substantially split up the U.S. West group in particular) would have accounted for the results better. Another contributing factor could be that the U.S. West had the lowest levels of lexicalizing of [ʌj], with just 3 percent of its raising participants indicating that *spider* rhymed with *writer*, compared to 34 percent in the U.S. North. To the extent that lexicalization of [ʌj] is related to [ʌj]~[aj] having the status of being partially contrastive, [ʌj]~[aj] is not partially contrastive in the U.S. West to the extent that it is in the U.S. North. However, the Canadians also had notably lower rates of [ʌj]-lexicalization than the U.S. North, despite similar accuracy in this experiment.

Therefore, while the lower prevalence of raisers in the U.S. West (and the lower prevalence of [ʌj]-lexicalization even among those who raise) might partially explain the reduced or absent partial contrast pattern in the accuracy results of the U.S. West group, there is probably another factor at play. One candidate for this is dialect exposure, specifically exposure to raised diphthong variants. It could be the case that encountering raised variants in the speech of other people (especially minimal pairs like *writer* and *rider*) produces or reinforces the partially contrastive status of [ʌj]~[aj], even in speakers who already have raised variants in their own speech. Thus, even the people who do raise in the U.S. West might be affected by the fact that they encounter raising less often in the speech of others in their own region. In addition to exposure to raised variants from people in their own region, there is also reason to believe that the U.S. North has a greater level of exposure than the U.S. West to Canadians and Canadian English. This could be expected based on the greater proximity between the U.S. North and Canada (taking into account the overall location of the regions but also the population centres), and there are figures that support this, including from U.S. travel to Canada.

Figure 6.6 shows the top 15 states of origin for travellers to Canada in 2014 by total number of trips, unadjusted for population (Statistics Canada, 2015). To get a more accurate perspective on how much time people from each state spend in Canada, the values in the figure indicate the number of nights (rather than the number of trips) adjusted by the state's population as of July 1, 2014 (U.S. Census Bureau, 2019).[3] While the top state for travel to Canada (by a significant margin) is Washington (724 nights in Canada per 1,000 state residents), which is in the U.S. West region, Washington also contributed a relatively small number of participants to the U.S. West sample in this experiment. The largest number of participants in the U.S. West sample came from California, which has the second lowest rate of travel to Canada (129 nights per 1,000 residents) and is noticeably lower than the rate of travel to Canada of the three most represented states in the U.S. North sample: New York (280 nights per 1,000 residents), Pennsylvania (175 nights), and Michigan (278 nights). By using the travel data for these 15 states and weighting it according to the state origins of the participants in this study, it can be estimated that the rate of visiting Canada is 208 nights per 1,000 residents in Experiment 5's U.S. West sample and 239 nights per 1,000 residents in the U.S. North sample. This is only a partial estimate, due to the lack of data for the states of 10 participants in the U.S. West (Arizona, New Mexico, Utah, Colorado, and Idaho) and for 1 participant in the U.S. North (New Hampshire). Those additional states in the U.S. West presumably pattern more closely with California's low rate of travel to Canada than Washington's high rate of travel to Canada, and so a more complete data set on travel would probably lower the number for the U.S. West sample.

Travel to Canada is just one measure that could index dialect exposure, and a more comprehensive comparison of exposure to Canadians and Canadian English among residents of the U.S. North and U.S. West could also include Canadian media, travel to the United States by Canadians, friend and family ties across the border, and dual citizenship. Nevertheless, travel to Canada does provide at least limited evidence that the U.S. North has greater exposure to Canadians and Canadian English than the U.S. West does.

The possible role that dialect exposure plays in the perceptual effects of partial contrast, and the related disconnect between perception and production (as the model based on region was better supported than the model based on production), should be investigated further, as it was covered in only one experiment in this dissertation (while the context-dependent discrimination effect, or the flap advantage, was the subject of a series of experiments). Future research on the cross-dialectal perception of Canadian Raising would ideally include a more detailed questionnaire for American participants in particular, including local questions (such as their impressions on whether people in their area pronounce *rider* and *writer*

[3] The states without values in this figure have a lower absolute number of trips to Canada, but this does not necessarily mean that they have a lower number of population-adjusted trips. It is possible that some smaller states not included on this map (such as Vermont) have a higher rate of travel to Canada than larger states that are included (such as Texas).

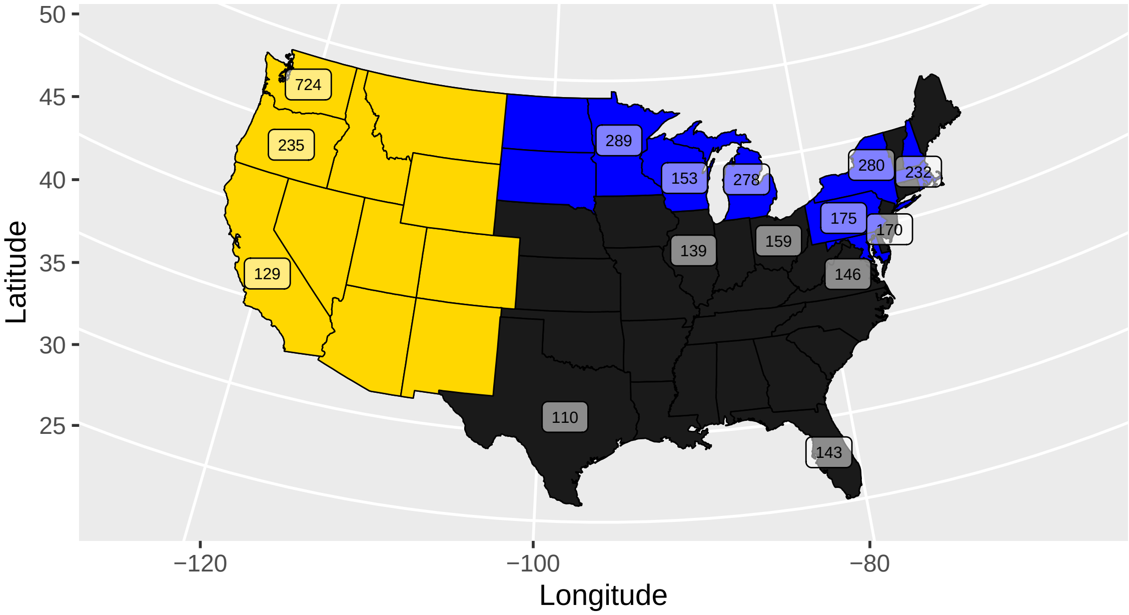

Figure 6.6: Top 15 states for travel to Canada by trips (2014), with values indicating number of nights spent in Canada per 1,000 state residents (gold indicates states in the U.S. West dialect region for the purposes of Experiment 5, and blue indicates states in the U.S. North dialect region)

differently, and whether they personally interact more with local or national media) and questions about level of exposure to Canadians and Canadian English (their level of travel to Canada, exposure to Canadian media, and family/friend ties in Canada). This could help clarify the role of variant or dialect exposure in partial contrast perception effects.

One final point to mention is that the discussion of group differences so far has focused on the overall performance, across the three environments (isolation, voiceless, and voiced). The isolation environment was included to test an additional hypothesis, and the voiced environment was included due to the finding in past experiments that the voiced environment provides listeners with an important, if not fully understood, source of input. However, among these three, only the voiceless environment is the naturally occurring environment for raised and non-raised diphthongs, and as such, a future experiment should perhaps focus on the voiceless environment alone (whether not including the other environments in the experiment, or not including them in the analysis).[4] Looking back to the accuracy results by participant region in Figure 6.2, focusing on the voiceless environment might have resulted in a three way distinction in performance on /aj/ where the Canadian group pulls ahead of the U.S. North group in performance (rather than being comparable to the U.S. North and only above the U.S. West). A version of that plot with

[4] Only including the voiceless environment would also allow a wider range of items to be used. The words used in this experiment were limited to those with no onset or an /h/ onset because other onsets would have caused undesirable formant transitions in the vowel that would remain after the onset was spliced off to make the isolation condition.

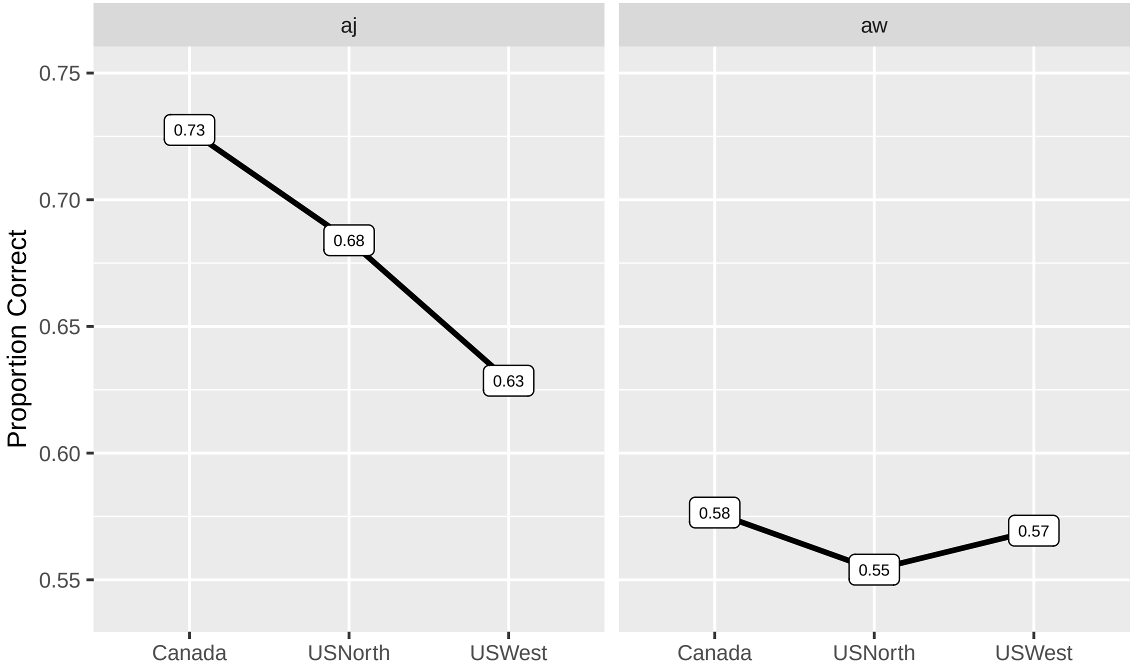

Figure 6.7: Discrimination of diphthong variants in Experiment 5 by participant region (voiceless context only)

only the voiceless environment has been reproduced here as Figure 6.7.

6.1.4.1.2 Response Time

The Canadians were 340 ms faster in their responses than the U.S. North, and 445 ms faster than the U.S. West. This is similar to the accuracy data in that the performance of a group is positively correlated with their "connection" to Canadian Raising (in production and/or exposure), which is highest for the Canadians, lowest for the U.S. West, and somewhere in the middle for the U.S. North. However, the key difference between the response time results and the accuracy data is that accuracy had an interaction between diphthong and region, with the U.S. West group performing similar to the other two groups on the /aw/ diphthong but worse on the /aj/ diphthong, which pointed towards a role of partial contrast in explaining that outcome variable. In the response time data, however, there was no such interaction between diphthong and group. There was not even an overall advantage for the /aj/ diphthong over the /aw/ diphthong on response time (participants were slightly slower on /aj/ trials at 1833 ms compared to 1820 ms), despite all of the previous findings in accuracy that participants (at least Canadian participants) have higher accuracy for discriminating raised and non-raised variants of /aj/ than for /aw/. This discrepancy between the response time results in this experiment and the accuracy results (of this experiment and past experiments) suggests that partial contrast might not play a role in these response time results.

This results in an interesting discrepancy. In accuracy, production of and/or exposure to Canadian Raising results in a partially contrastive status for [ʌj]~[aj] (but not [ʌw]~[aw]) in an individual's phonology, which gives them a discrimination boost for [ʌj]~[aj] as a partial contrast effect even in the voiceless environment. In response time, production of and/or exposure to Canadian Raising results in faster discrimination for both [ʌj]~[aj] and [ʌw]~[aw], without apparently being related to partial contrast. There are many open questions here, including why there is not a diphthong difference at least for the two American groups. If Canadians have the fastest overall performance because of their production of (and exposure to) Canadian Raising, would not the Americans have faster performance on /aj/ than /aw/ due to the greater prevalence of /aj/ raising than /aw/ raising that has been found in the United States? Does this indicate that exposure to raised and non-raised variants of /aj/ carries over and also speeds up discrimination of /aw/ variants, or does this actually indicate that /aw/ raising is more common in the United States than has been previously found (just like the survey results in this experiment found greater levels of /aj/ raising in the U.S. West than would be expected based on Labov et al., 2006)? Nevertheless, the finding of a complex interplay between production, dialect exposure, and perception, affecting response time and accuracy differently (seemingly involving partial contrast for accuracy but not response time) is a compelling first result for the study of cross-dialectal perception of Canadian Raising. Although response time has been used as an outcome variable for studying partial contrast before (e.g., Hume and Johnson, 2003), this result shows that partial contrast effects might be more likely to show up in accuracy data.

6.1.4.1.3 Linguistic Stereotypes and Speech Perception

Setting aside the nuances of the two different outcome variables discussed above (and the differences between grouping participants by production or by region, which is related to both production and exposure), there is one broader finding that deserves special attention, namely that a greater level of Canadian Raising (in production and/or exposure) was associated with faster and more accurate responses. Overall, the highest accuracy and speed was found from the Canadians, followed by the U.S. North, and then the U.S. West, contrary to the initial hypothesis, which was based on the idea that raised diphthongs are more marked or noticeable for non-raisers than non-raised diphthongs are for raisers, as suggested by the apparent fact that raising plays a much more prominent role in stereotypes of dialects than non-raising does. The lack of support for this hypothesis in Experiment 5 is consistent with Niedzielski (1999), who found that speakers of American English from Detroit had their perception of a speaker's vowels highly influenced by the speaker's (perceived) nationality. Participants listened to the speech of a fellow resident of Detroit, and they were asked to choose from a set of resynthesized vowels which pronunciations best matched the speaker that they heard. When they were told that the speaker was from Canada, participants picked more raised tokens compared to if they were told that the speaker was from Michigan. The results of this experiment, alongside the results of Niedzielski (1999), provide evidence that noticing stereotypical dialectal features in the speech of others is dependent to a large extent on expecting those features based on pre-existing knowledge of the speaker's dialect, and is not necessarily driven by any heightened acuity or sensitivity to those features in speech. This finding is strengthened by two other observations, namely the phonetic inaccuracy of the *oot and aboot* stereotype (which is inaccurate because raised diphthongs are still diphthongs) and the fact that raising plays such a prominent role in stereotypes of Canadian English on the part of Americans despite being (at least for /aj/) very common in the United States as well.

6.1.4.2 Context Effects in Perception

The second hypothesis that was tested is that speakers of Canadian English will have an easier time discriminating between diphthong variants when they are presented in isolation, compared to being presented in the allophonic (voiceless) context. This was not found in the results. The response time data did not have any statistically significant effect or interaction involving environment, but the accuracy data actually found a significant interaction in the opposite direction to what would be predicted by the hypothesis. The Canadian group's performance in the isolation condition, relative to the voiceless condition, was worse than that of the other two groups, to the point of a statistically significant disadvantage in the isolation condition compared to the voiceless condition for the Canadians. However, as seen in Figure 6.2 (and the lack of significant difference between the voiceless and voiced environments for the Canadians in /aj/), this effect should be understood as being driven more by high performance for the Canadians in the voiceless environment than by low performance in the isolation context, which was previously explained as probably being a partial contrast effect (a boost for discrimination of partially contrastive segments not just in the specific contrastive environment but also in other environments, presumably specifically licit environments). Still, this does mean that there is no discrimination boost experienced by Canadians for taking the raised and non-raised variants out of the allophonic environment and presenting them in isolation.

6.1.4.3 Dialect Survey Data

In addition to the psycholinguistic data collected, this experiment also collected self-reported production data from the three regional groups (Canada, U.S. North, and U.S. West) on two topics: the presence of Canadian Raising of /aj/, and (among those who self-report raising of /aj/) the lexicalization of [ʌj] in *spider*. The self-reported raising results found the least raising in the U.S. West. However, raisers still made up a majority of the U.S. West sample (59 percent), which conflicts with the findings of Labov et al. (2006) that the U.S. West generally does not exhibit Canadian Raising. While it is reasonable to prefer acoustic analysis over self-reported data, there are reasons to take this self-reported data seriously, namely that the self-reported production results from the Canadian group closely lined up with the acoustic analysis of Canadians performed by Boberg (2008), and because Labov et al. (2006) performed acoustic analysis on recordings from a telephone survey carried out in the years 1992–1999, and it is possible that raising has become more prevalent in the U.S. West in the 20 or more years that have passed since those recordings.

The [ʌj]-lexicalization results found a significant effect of region, with the most [ʌj]-lexicalization found in the U.S. North (34 percent of participants with raising identified pronouncing *spider* with [ʌj]), followed by Canada (13 percent of raisers), and then finally the U.S. West (3 percent of raisers). It should be noted that these absolute rates might be underestimations of [ʌj]-lexicalization, because participants might be influenced by the spelling of the words (making them more likely to indicate that *spider* rhymes with *rider* and giving the impression of them lacking [ʌj]-lexicalization), but this effect would presumably not alter the differences between regions in rates of [ʌj]-lexicalization. As for the reasons behind the regional differences here, the greater prevalence of [ʌj]-lexicalization in the U.S. North than the U.S. West could be explained by raising being a newer phenomenon in the U.S. West, and lexicalization taking additional time to progress. The relatively low rate of [ʌj]-lexicalization in Canada compared to the U.S. North cannot be explained in the same way, because raising in Canada (specifically Ontario, where a large portion of the Canadian sample is from) has been documented as far back as people born in 1860 (Chambers, 2006a). It is possible that [ʌj]-lexicalization is being influenced and driven by some other factor in the U.S. North, such as the Northern Cities Vowel Shift, but the details are unclear.

The survey findings on [ʌj]-lexicalization should be considered in light of Hall (2005), who finds unexpected uses of [ʌj] and [aj] that conflict with the pattern of Canadian Raising ([ʌj] in non-raising environments and [aj] in raising environments) among speakers in Rural Ontario. She argues that these apparent conflicts with Canadian Raising can be explained by the lexical neighbourhood effect and the *preceding* consonant, which means that salient words with or without raising affect the raising status of other words that share the same onset. If this is the case for *spider*, it would mean that a salient word with [spʌj] (such as [ˈspʌjt] *spite* or [ˈspʌjs] *spice*) influences the pronunciation of other words with an /sp/ onset, such as *spider*, to be pronounced with a raised vowel. This lexical neighbourhood effect is a possible alternative explanation for this unexpected [ʌj] in *spider* that might not require saying that [ʌj] is lexicalizing in some dialects. However, there are a few reasons why the lexical neighbourhood effect might not replace [ʌj]-lexicalization as an explanation for the unexpected [ʌj] in *spider*. The first is that *spider* is not an infrequent word; the log-transformed frequency per million words of *spider* in the SUBTLEX corpus is 2.71, which is 1.22 standard deviations above the mean log-frequency of 1.66 in that corpus (Balota et al., 2007). That can be compared to 1.08 SDs above the mean for *spite* or 0.90 SDs above the mean for *spice*, which means that these are not obvious candidates for influencing *spider* (in addition, there are frequent words with [spaj] like *spy*, which is 1.57 SDs above the mean log-frequency). The second reason is that it is not clear why the lexical neighbourhood effect would impact dialects, particularly high raising dialects like Canada and the U.S. North, differently. Canadian English, the dialect that her study on the lexical neighbourhood effect was focused on, actually had a noticeably lower rate of [ʌj]-lexicalization than the U.S. North. As mentioned above, the [ʌj]-lexicalization explanation also does not have a clear explanation for the dialect differences (why might [ʌj]-lexicalization be more common in the U.S. North than Canada?), but at the very least it is no worse than the lexical neighbourhood explanation here. And finally, it is not clear the extent to which the lexical neighbourhood effect is an alternative explanation to lexicalization for unexpected instances of [ʌj] as opposed to a separate phenomenon or even a factor that contributes to [ʌj]-lexicalization. More work on unexpected instances of [ʌj] and [aj], particularly in the U.S. North (such as a detailed study like that of Hall, 2005, but for that region), would help clarify this.

Chapter 7

Discussion

This chapter reviews the primary findings of the investigation into the perception of Canadian Raising undertaken in this dissertation. The findings cover four main topics: (partial) contrast, [ʌj]-lexicalization and possible emergence of /ʌj/ as a new phoneme, cross-dialectal perception, and online data collection.

7.1 (Partial) Contrast

This dissertation was primarily an investigation into partial contrast and its impact on speech perception. Partial contrast is a term referring to phonological relationships that are intermediate or variable between contrast and allophony; for example, [r] and [ɾ] in Spanish have minimal pairs intervocalically (e.g., *carro* "car" and *caro* "expensive"), suggesting contrast, but are otherwise predictably distributed, suggesting allophony (Hall, 2013). The case of partial contrast that has been the focus of this dissertation is Canadian Raising. Although canonically an allophonic alternation, with raised diphthongs [ʌj, ʌw] occurring before voiceless consonants and non-raised diphthongs [aj, aw] occurring elsewhere, North American /t/–/d/ flapping creates the possibility of minimal pairs before a flap [ɾ], perhaps most notably [ˈɹʌjɾɪŋ] *writing* (underlying /t/) and [ˈɹajɾɪŋ] *riding* (underlying /d/). Although raised and non-raised diphthongs are overall quite predictable when taking into account all possible environments, in the particular environment of a flap they are distinctly unpredictable, at least on the surface (Hall, 2012).

The investigation into partial contrast and speech perception undertaken by this dissertation was grounded in past research into behavioural consequences of contrast in the traditional binary sense (where sounds are either seen as contrastive or non-contrastive), particularly research showing that native contrasts are easier to discriminate than allophones (Whalen et al., 1997; Boomershine et al., 2008; Barrios et al., 2016) and non-native contrasts (e.g., Goto, 1971; Werker et al., 1981; Sundara et al., 2006), at least when those non-native contrasts are mapped onto the same native phoneme (Best et al., 2001). This dissertation was also grounded in the more recent body of research that has started looking into partial contrast itself, which has shown partial contrast effects that follow two general patterns. This literature has found what could be called general partial contrast effects, where participants respond differently to partially contrastive sound pairs than to fully contrastive or fully allophonic sound pairs (Stevenson and Zamuner, 2017; Hume and Johnson, 2003); it has also found context-based partial contrast effects, where participants respond differently to the same partially contrastive sound pair in contrastive environments than in non-contrastive (allophonic or neutralizing) environments (Celata, 2008; Murphy et al., 2016).

Based on this work, the dissertation started with the hypothesis of a context-based partial contrast effect for Canadian Raising in discrimination. Specifically, the prediction was that there would be better discrimination of raised and non-raised vowels ([ʌj]~[aj] and [ʌw]~[aw]) in the flap (contrastive) environment than in the voiceless (allophonic) or the voiced (non-allophonic but also non-contrastive) environments. This effect was investigated over the course of the five primary experiments in this dissertation (in Chapters 3, 4, and 6) and the five additional secondary experiments (in Chapter 5), which produced five main findings about partial contrast as a phenomenon and how it applies to Canadian Raising that will be reviewed here.

The first main finding of this dissertation is that Canadian Raising does exhibit a context-dependent partial contrast discrimination effect, comparable to the context-based partial contrast effects found in Murphy et al. (2016) and especially in Celata (2008).[1] As first shown in Experiment 1, and clarified in later experiments, discrimination of raised and non-raised vowels differs between environments, based on the contrastive status of raised and non-raised vowels in those environments (among other factors). This finding provides behavioural evidence in favour of a view of contrast as gradient rather than binary, which is to say that contrast should be understood as involving not only the options of "contrastive" and "non-contrastive" but also a range of possible intermediate options (Hall, 2013). This is not to say that a binary conception of contrast is never useful, but rather that our understanding of contrast should not be limited to only the binary view (see Hall and Hall, 2016, on how gradient and binary conceptions of contrast can coexist). If discriminability, one of the behavioural dimensions that has been found to differ between

[1] These both involved identification tasks. Celata (2008) was more comparable to a discrimination task because it tested speed and accuracy of identification, while Murphy et al. (2016) tested perceptual preferences for ambiguous sounds on a continuum.

contrastive and non-contrastive sound pairs, can vary for the same sound pair based on the hypothesized different contrastive status of that sound pair in two different environments, then that provides evidence for that sound pair actually having a different contrastive status in those two different environments (in a sense, varying between contrast and allophony).

However, one important detail of this context-dependent discrimination effect result was that it was found for only one of the diphthongs involved in Canadian Raising: /aj/. This provides evidence in the specific case of Canadian English that only raised and non-raised variants of /aj/—and not the equivalent variants of /aw/—have the status of partially contrastive. Although [ʌw]~[aw] minimal pairs in the flap environment are in principle possible just like [ʌj]~[aj] minimal pairs, actual examples rely on uncommon or obscure words (e.g., the *clouting/clouding* or *doubting/Dowding* minimal pairs for [ʌw]~[aw], compared to *writing/riding* and *sighting/siding* for [ʌj]~[aj]) and as a result they might not be recognized as real word minimal pairs by regular listeners. The lack of behavioural evidence for partial contrast in /aw/ (a lack of context-dependent discrimination effect) suggests that the theoretical possibility of having minimal pairs for [ʌw]~[aw] does not actually make that sound pair partially contrastive. Presumably, if minimal pairs did exist for /aw/ and were recognizable to participants, the behavioural results would have been different.

Alternatively, rather than saying that [ʌw]~[aw] is not a case of partial contrast, it might be useful to distinguish between two different types or definitions of partial contrast. One type of partial contrast could refer to whether (for a normally allophonic sound pair) minimal pairs *can* be made in one environment, which is indeed the case for [ʌw]~[aw]. Another type of partial contrast could refer to whether minimal pairs actually exist and are recognizable by speakers, which does not apply to [ʌw]~[aw] (at least based on the present behavioural evidence). This distinction between partial contrast in these two senses might provide insight into the mechanisms behind the (context-specific) effect of partial contrast on perception. Discrimination accuracy appears to be driven not just by an abstract knowledge of whether minimal pairs are possible, but rather by actual experience with minimal pairs. It appears that listeners have heightened accuracy for [ʌj]~[aj] before a flap (but no heightened accuracy for [ʌw]~[aw] before a flap) because they have experience with [ʌj]~[aj] minimal pairs before a flap (like *writing/riding*) but no [ʌw]~[aw] minimal pairs before a flap, even though they technically could be made.

In addition to finding that the context-dependent discrimination effect applied only to /aj/ and not /aw/, it was also found that only the flap and voiceless environments (but not the voiced environment) produced discrimination results according to their contrastive status; the voiced environment (which is not contrastive for raised and non-raised diphthongs) exhibited higher rates of discrimination, more like the contrastive flap environment than the non-contrastive voiceless environment. However, this is not necessarily relevant for partial contrast because the voiced environment differs from the other two environments in important ways. The voiced environment has a significantly longer vowel duration, and raised diphthong variants do not actually canonically occur in the voiced environment in any relevant dialects (dialects that Canadians would likely have significant exposure to). Both of these additional factors could plausibly affect the discriminability of diphthong variants in the voiced environment.

The second main finding of this dissertation is that the context-dependent discrimination effect found in Experiment 1 can be broken down into two components: an inherent discrimination advantage for partially contrastive segments in the contrastive environment over the non-contrastive environment, as well as an additional discrimination advantage for partially contrastive segments when they create minimal pairs (which can only happen in the contrastive environment). The "inherent" effect of environment can be seen in the results of Experiment 2, which tested discrimination of raised and non-raised variants in non-words, meaning that the stronger minimal pairs for [ʌj]~[aj] (like *writing/riding*) and the weaker minimal pairs for [ʌw]~[aw] (like *clouting/clouding*) from Experiment 1 were both replaced with items that were clearly not minimal pairs at all (at least in the sense of minimal pairs involving real words with different meanings). Even in the complete absence of lexicality, Experiment 2 still found a difference between diphthongs—the /aj/ diphthong exhibited a pattern that was closer to the flap advantage from Experiment 1 than the pattern found for /aw/. This partial contrast effect in the absence of lexical meaning compares to Celata (2008) and Murphy et al. (2016), who both used non-words.

However, the results from Experiment 2 were not as strong as those from Experiment 1, suggesting that even if a difference between /aj/ and /aw/ can be found in the absence of the strong minimal pairs for /aj/, those strong minimal pairs for /aj/ might still have played a role in the findings from Experiment 1. Experiments 3 and 4 (focusing on words with the /aj/ diphthong) investigated the effect of lexicality; Experiment 3 showed more of a flap advantage for a flap condition where [ʌj]~[aj] create different words (like *writer/rider*) than a flap condition where they do not (like *fighter*), while Experiment 4 found that the same contrast (*writing/riding*) is discriminated better when preceded by a word (like *maybe*) that allows both interpretations than when preceded by a word (like *book*) that strongly encourages one meaning over the other. Lexical effects have received a considerable amount of attention in the speech perception literature; for example, phonological categories can be perceived or filled in based on higher-level lexical knowledge and expectations when a sound is ambiguous (as in the Ganong effect: Ganong, 1980; Pitt and

Samuel, 1993) or has been replaced by noise (as in the phoneme restoration effect: Warren, 1970; Samuel, 1981). However, the lexical effect in this dissertation (which appears to be one component of the overall partial contrast effect in Experiment 1) is different from those findings; it involves better discrimination for two sounds when they create minimal pairs compared to when they do not. This is reminiscent of some past studies that have found better performance for real words over non-words on various tasks, such as higher correct responses on an AX discrimination task for real words (like *loss/lot*) than non-words (like *voss/vot*) in children with dyspraxia and healthy controls (Bridgeman and Snowling, 1988) as well as higher correct responses on a word repetition task for real words (e.g., *lemon*) compared to non-words (e.g., *fepon*) in elderly patients with Alzheimer's disease and healthy controls (Glosser et al., 1997). These two components of the overall partial contrast effect—the inherent discrimination difference between the contrastive and non-contrastive environments, and the additional discrimination boost for minimal pairs—were weaker on their own (in Experiments 2, 3, and 4), but together they can help explain the stronger partial contrast effects in Experiment 1.

The third main finding of this dissertation is that this context-dependent discrimination effect is heavily dependent on the listener having a certain quantity and/or quality of exposure to the voices of the speakers. Experiments 1–4, which all found some effect of partial contrast, were additionally tested (in Chapter 5 as Experiments 1b, 1c, 2b, 3b, and 4b) with the removal of the voiced condition in the experiment (words like *ride*) on new samples of listeners. The removal of the voiced environment tokens from the experience of the participants affected the relationship between the voiceless and flap environments, with the partial contrast effects discussed above failing to replicate in the absence of the voiced tokens in every single case but one. That exception was the sample of participants in Experiment 4b who were recruited personally from the friend groups of the individuals whose voices were recorded for the exerpiment. This sample of participants in Experiment 4b (who had previous exposure to the voices of the speakers) exhibited the same partial contrast effect that was found in Experiment 4 (which included the voiced condition), while the other sample of participants in Experiment 4b without previous exposure to the speakers did not.

It is possible that the presence of the voiced condition mattered simply because it added to the overall degree of exposure to the speakers' voices. Experiment 1, for example, had a total of 192 trials, with 64 in the voiceless environment (half for /aj/ and half for /aw/), 64 in the flap condition, and 64 in the voiced condition. Removing the voiced tokens from the experiment removed one third of the exposure that the listeners had to the voices of the speakers. On the other hand, it is also possible that the voiced environment provided a particular type of exposure to the speakers' voices (such as durationally longer vowels) that gave special benefit to the listeners in creating a model of each speaker's vowel space and phonological patterning in a shorter time frame than might otherwise be necessary. As discussed in Chapter 5, one promising idea is that the longer duration vowels in the voiced condition are easier to process (evidenced by the higher discrimination in the voiced condition), and as such they provide special benefit for the listeners in creating models of the speakers' vowel spaces.

One question that could be raised is why the effect of partial contrast in perception is so dependent on the experimental conditions or the level of exposure that the listeners have to the speakers' voices. It is possible that the elusive nature of this context-based partial contrast effect is related to the fact that it involves vowels rather than consonants. As mentioned in the introduction, past research has found that consonants tend to be perceived more categorically than vowels and tones (i.e., there is a greater effect of category boundary or contrast on perception). It could be that the context-based partial contrast effect (difference between contrastive and non-contrastive environment) is smaller for this vowel phenomenon than it would be for a comparable consonant phenomenon, making it more susceptible to disappearing unless under ideal conditions.

Regardless of whether this effect occurs by means of providing a higher degree of exposure to the speakers' voices or by providing a particular kind of exposure, and regardless of whether consonants would behave differently, this finding is reminiscent of past results on talker-specific effects where participants respond differently in tasks such as word recognition and word shadowing depending on whether the stimuli they encounter are produced by the same speaker or different ones (Goldinger, 1996; Luce and Lyons, 1998; Goldinger, 1998) or respond differently when listening to familiar talkers than unfamiliar ones (the Familiar Talker Advantage: Souza et al., 2013; Case et al., 2018b,a). These effects provide evidence that idiosyncratic aspects of speech (such as voice details) are stored in memory and affect later perception, rather than being simply filtered out as noise. This finding has been used in support of episodic or exemplar models of speech perception and the lexicon, which are most tenable when they also allow a role for abstract lexical representations alongside this episodic detail in long-term memory (Luce and Lyons, 1998; McQueen et al., 2006; Monahan, 2009). The effect of input/exposure found in this dissertation supports the view that voice details are stored in memory and have an influence on later perception. This effect occurred over a shorter time-frame in the case of the partial contrast effect being facilitated by the inclusion of voiced tokens in an experiment, because the experiments took on average 15 minutes to complete. However, this would have to involve voice details stored in memory over a longer time-frame in the case of the one sample of participants in Experiment 4b that had exposure to the speakers' voices. Although they participated anonymously, it can be assumed (in part because of their geographical distribution) that many or most of them had not heard the voice of the speakers in the days preceding their participation in the experiment.

The fourth main finding of this dissertation is that, in addition to the context-based partial contrast effects (where participants respond differently to the same partially contrastive sound pair in contrastive environments than in non-contrastive environments), Canadian Raising also involves what could be called a general partial contrast effect, where participants respond differently to partially contrastive sound pairs than to fully contrastive sound pairs or fully allophonic sound pairs, as in Hume and Johnson (2003) and Stevenson and Zamuner (2017). In other words, in addition to the fact that discrimination accuracy for [ʌj]~[aj] (but not [ʌw]~[aw]) is higher in the flap environment than the voiceless environment, discrimination of [ʌj]~[aj] is better than [ʌw]~[aw] even in the voiceless environment. This can partly be explained by the greater phonetic distance between raised and non-raised variants of /aj/ than /aw/, as found in the phonetic analysis in Chapter 2. However, Experiment 5 found a regional difference suggesting that partial contrast also plays a role in this, specifically that part of the advantage for /aj/ over /aw/ even in the voiceless environment is a result of [ʌj]~[aj] but not [ʌw]~[aw] being partially contrastive. The regional difference was that the U.S. West (but not the U.S. North) differed from Canada in having a smaller accuracy difference between diphthongs. The U.S. West also differs from the other two regions in its lower prevalence of raising, which would plausibly result in [ʌj]~[aj] not having the status of partially contrastive in this region to the same extent as the other regions. However, there was evidence that this lower prevalence of raising in the U.S. West should be understood not solely in terms of lower rates of raising by speakers in that region, but also in terms of lower rates of dialectal exposure to raised diphthong variants. Thus, broadly, input or exposure appears to matter for this general partial contrast effect, in addition to input mattering for the context-dependent partial contrast effect as previously discussed.

Finally, the fifth main finding of this dissertation is that partial contrast effects are not manifested equally in response time and accuracy. Response time was not used as a measure for investigations into the context-dependent partial contrast effect (Experiments 1–4 and their variations, because stimulus duration varied between the contexts), but when response time was included in Experiment 5, the general effect of partial contrast was only found for accuracy and not response time. To compare to past work on partial contrast, Stevenson and Zamuner (2017) found a general effect of partial contrast in both accuracy and response time, although the effect was clearer and more consistent for accuracy; Celata (2008) found a context-based effect of partial contrast in both accuracy and response time, and Hume and Johnson (2003) found a general effect of partial contrast in response time (accuracy results were not presented). To compare to past work on contrast versus allophony and discrimination, Whalen et al. (1997) found a benefit for contrast over allophony in accuracy, Boomershine et al. (2008) found one in response time, and Barrios et al. (2016) replicated the finding of Boomershine et al. (2008) using accuracy. Overall, there is no clear precedent from the past literature on the perception of partial contrast or the discrimination of contrast versus allophony that would give rise to a clear prediction that only accuracy but not response time would be influenced by (partial) contrast, but the present results do suggest that the outcome variables can be affected differently.

7.1.1 Future Work on Partial Contrast

These results raise various issues that should be considered by future studies into partial contrast. Most important is the finding that partial contrast effects are dependent on two things: the outcome variable used (accuracy appears to show partial contrast effects more than response time), and the quantity/quality of exposure that listeners have to the speakers' voices. Future perception studies of partial contrast should take these findings into account because it is possible with any particular investigation into a partial contrast phenomenon that an effect of partial contrast on discrimination exists and could be found, but is not found because the experiment and its analysis were planned around response time only, or because the listeners were not adequately familiar with the voices of the speakers used in the stimuli. The resulting recommendation regarding the outcome variable is straighforward: look at accuracy, or both accuracy and response time. Unfortunately, it is more difficult to make specific recommendations regarding exposure to speakers' voices because the nature of that effect remains mysterious.

It is possible that the existence of the voiced environment tokens in Experiments 1–4 simply provided a greater quantity of exposure to raised and non-raised vowel tokens produced by the speakers, but it is also possible that there was something special about the voiced environment, such as its longer vowel duration (giving participants a longer window to process the vowels). It is also not clear whether this effect is specific to vowels, or whether a similar effect of input would also be found for a partially contrastive set of consonants. These questions surrounding the importance of input/exposure in partial contrast could be fruitful areas of future research into partial contrast, alongside the questions surrounding the difference between accuracy and response time (why might partial contrast manifest differently in accuracy and response time?).

There also remains the task of expanding the study of how partial contrast influences perception to a wider range of partial contrast phenomena across languages. For example, the Spanish [r]~[ɾ] alternation, which has been referenced here as another example of partial contrast, would potentially be a good candidate. This dissertation has, alongside Hume and Johnson (2003), Celata (2008), Murphy et al. (2016), and Stevenson and Zamuner (2017) documented a variety of different partial contrast effects, in Mandarin, Italian, Canadian French, and Canadian English. There are general partial contrast effects, where participant respond differently to partially contrastive sound pairs than to fully contrastive or fully allophonic sound pairs (Hume and Johnson, 2003; Stevenson and Zamuner, 2017), and there are context-based partial contrast effects, where participants respond differently to partially contrastive sound pairs in contrastive environments than in non-contrastive (allophonic or neutralizing) environments (Celata, 2008; Murphy et al., 2016), with this dissertation primarily providing more results on the latter kind of partial contrast effect. With both of these effects, contrast is associated with faster and/or more accurate discrimination. These findings provide a blueprint for investigation of partial contrast effects in other phenomena (such as the Spanish alternation). Study into other phenomena would clarify this typology of partial contrast effects on perception, and help understand the differences and similarities between different types of partial contrast phenomena. For example, given findings that consonants produce categorical perception effects more strongly than vowel segments or tones, to what extent do the perceptual consequences of partial contrast differ between consonants, vowels, and tones? Are there differences between partial contrast phenomena that are primary contrastive (closer to the contrast end of the spectrum, such as a contrast that gets neutralized in certain limited environments) and those that are primarily allophonic (closer to the other end of the spectrum, such as an allophonic alternation like Canadian Raising that has a limited number of minimal pairs)?

7.2 Perceptual Basis for the Emergence of a New Phoneme

One of the earliest reports of Canadian Raising, Joos (1942), speculated that the allophonic alternation might result in the emergence of two new raised phonemes: /ʌj/ and /ʌw/. There has been some evidence that [ʌj] (but not [ʌw]) is lexicalizing and emerging as a new phoneme, although as of yet it appears to be limited to certain speakers of certain dialects. Vance (1987) reports on speakers from the Northern United States who have [ʌj] in *spider* and *cider* (which cannot be the Canadian Raising allophonic alternation as traditionally described because the flap is not underlyingly a /t/ to trigger raising) and even an [ʌj]~[aj]

minimal pair in *idle/idol.* Hall (2005) also finds a degree of unexpected use of [ʌj] in non-raising environments among speakers in Rural Ontario (as well as unexpected use of [aj] in raising environments), although she argues that many of these cases can be explained by the lexical neighbourhood effect and the *preceding* consonant; salient words with or without raising affect the raising status of other words that share the same onset (e.g., the salient word *meningitis* with allophonic raising causes unexpected raising after /dʒ/ in words like *gigantic* and *angina*). This might be seen as an alternative explanation to the possibility that [ʌj] is lexicalizing as a new phoneme—or it might be one of the contributing factors to *why* [ʌj] might be lexicalizing as a new phoneme (and for why it is happening in those particular words). Further evidence for the (dialect-specific) lexicalization of [ʌj] comes from the questionnaire in this dissertation's Experiment 5, which asked participants whether *rider* and *writer* sound the same or different in their own natural pronunciation (a response of different suggests that they have Canadian Raising for /aj/) and asked those who self-report having raising of /aj/ whether *spider* rhymes with *rider* or with *writer*. Responding with *writer* indicates that they pronounce *spider* with an [ʌj]. A full 34 percent of participants from the Northern United States (see Figure 6.1 for the boundaries used and the distribution of participants by state) who raise reported this lexicalized [ʌj], compared to 13 percent of participants from Canada and just 3 percent of participants (one speaker) from the Western United States. While this is self-reported evidence from just one word, it does appear to show lexicalized [ʌj] for a large minority of speakers in the Northern United States, as well as large regional differences. These figures might even underestimate the prevalence of [ʌj]-lexicalization, because based on spelling we might expect participants to be biased in favour of answering that *spider* rhymes with *rider* (which would be interpreted as no [ʌj]-lexicalization), regardless of their pronunciation.

In addition to providing evidence that lexicalization of [ʌj] is happening (particularly in the Northern United States, near the Great Lakes), this dissertation also provides evidence for why it is happening—specifically, why it is happening for /aj/ but not /aw/. The phonetic analysis in Chapter 2 of the recordings made for the stimuli found a larger phonetic difference between raised and non-raised variants of /aj/ than /aw/, which would predict that raised and non-raised variants are easier for listeners to discriminate for /aj/. If discrimination really is easier for [ʌj]~[aj] than [ʌw]~[aw], it would seem that a phonologigical contrast is more likely to emerge for [ʌj]~[aj] than [ʌw]~[aw]. The experiments of this dissertation found clear and consistent evidence that discriminability is higher for raised and non-raised variants of /aj/ than /aw/, with a discrimination advantage for /aj/ (of approximately 5 to 10 percentage points) found in every experiment that tested discrimination of both diphthongs: Experiments 1, 2, 1b, 2b, and 5. Additionally, Experiment 5 found that this discrimination difference is particularly large for Canada and the U.S. North (compared to the U.S. West), suggesting that it is driven not just by phonetic factors but also by phonological ones, specifically the status of [ʌj]~[aj] as partially contrastive (this is the general effect of partial contrast discussed previously). On top of this, there is an additional discrimination advantage for [ʌj]~[aj] in the flap environment, which is the "flap advantage" (context-based discrimination effect) that was the main focus of this dissertation. This flap advantage for /aj/ was found to be present even for non-words (Experiment 2), although it also appears to be related to minimal pairs and the lexicality of the words used (Experiments 3 and 4). Although this flap advantage for /aj/ was found less strongly and less consistently than the overall advantage for /aj/ (see Chapter 5 on the conditions under which this flap advantage is not found), the particularly high discrimination of [ʌj]~[aj] in the flap environment could contribute to the prevalence of lexicalized [ʌj] in the flap environment. Particularly strong discrimination of [ʌj]~[aj] before a flap would appear to make the *idle/idol* minimal pair more likely than a minimal pair before a [t].[2]

There are other additional factors that could also help explain the reason for the apparent lexicalization of [ʌj] but not [ʌw], including the fact that /aj/ is overall a more common vowel than /aw/ (3.6 times more common in type frequency and 4.1 times more common in token frequency, according to the SUBTLEX corpus accessed through IPhOD database Brysbaert and New, 2009; Vaden et al., 2009) and the (possibly related) fact that more dialects have allophonic raising of /aj/ than allophonic raising of /aw/, at least in the United States (Labov et al., 2006). If lexicalization of the raised variant depends on the existence of the raised variant as an allophone then lexicalization of [ʌw] is less expected in the United States, although it is not less expected in Canada, where allophonic raising of /aw/ is as common (Boberg, 2008). However, this raises additional questions. Is lexicalization of [ʌj] less common in Canada than the Northern United States, as suggested by the questionnaire results in Experiment 5? If so, why would lexicalization of [ʌj] be more likely to emerge in the Northern United States than in Canada? Finally, why is Canadian raising more common for /aj/ than /aw/ in the United States, when they are approximately just as common in Canada?

Setting aside these questions for further research to summarize the present findings, there are at least five reasons (some of which are related to each other) for the apparent lexicalization of [ʌj] but not [ʌw], based on findings from other sources (reasons 1 and 2) and findings from this dissertation (reasons 3 to 5):

[2] The experiments in this dissertation also found elevated discrimination in the voiced environment. Although it was explained as likely a result of the longer duration in the voiced environment (and thus not theoretically interesting for the question of partial contrast), this discrimination advantage raises the expectation of a minimal pair like *idle/idol* before a [d] or other voiced consonant.

1. /aj/ is a more common vowel than /aw/ (Brysbaert and New, 2009; Vaden et al., 2009).
2. Raising of /aj/ is more widespread than raising of /aw/, at least in American English (Labov et al., 2006).
3. There is a greater phonetic difference for [ʌj]~[aj] than [ʌw]~[aw], leading to the [ʌj]~[aj] difference being easier to distinguish (based on the phonetic analysis in Chapter 2 and in Hall, 2015, and based on discrimination results in all experiments that tested both diphthongs).
4. There is additionally better discrimination for [ʌj]~[aj] in raising dialects based on the fact that [ʌj]~[aj] has the status of partially contrastive while [ʌw]~[aw] does not (as a result of, or at least demonstrated by, [ʌj]~[aj] having much more common or recognizable minimal pairs like *writing/riding*). This is the general effect of partial contrast, found in Experiment 5.
5. Also as a result of its partially contrastive status, in raising dialects [ʌj]~[aj] has an additional advantage for discrimination in the environment where it is contrastive: the flap environment. This is the context-based effect of partial contrast, found in Experiment 1 and clarified in later experiments.

7.3 The Effect of Dialectal Stereotypes and Exposure on Perception

Canadian Raising, especially for /aw/, is a stereotypical and frequently remarked on identifier of Canadian English that is often exaggerated as *oot and aboot* (Chambers, 1973; Boberg, 2008). On the other hand, with the exception of experiences almost 80 years ago reported in Joos (1942) ("if I use a low diphthong before a fortis consonant [...] the Canadian listener immediately accuses me of drawling", p. 142), non-raising (of either /aj/ or /aw/) does not appear to be a prominent feature of American English that is remarked on or stereotyped by Canadian English speakers. This leads to the impression that raised diphthongs are more marked or noticeable for non-raisers than non-raised diphthongs are for raisers. Based on this, it was hypothesized in Experiment 5—the only experiment to test speakers of American English in addition to speakers of Canadian English—that the existence of raising for an individual (or, stated in dialect terms, a greater prevalence of raising in a dialect) would be associated with lower discriminability for raised and non-raised diphthongs, while individuals without raising (or dialects with less raising) would better discriminate between the variants.

Experiment 5 tested speakers of Canadian English and speakers of American English from two regions: the U.S. North (around the Great Lakes) and U.S. West (the mainland United States as far east as Monatana, Wyoming, Colorado, and New Mexico). Based on Labov et al. (2006), raising of /aj/ (but not /aw/) is expected in the U.S. North while no raising is expected in the U.S. West, and thus the U.S. North was expected to be intermediate between Canada and the U.S. West in terms of performance. A short dialect survey at the beginning of Experiment 5 gathered data on participant /aj/ raising ("In your most natural pronunciation, do the words *rider* and *writer* sound the same or different?"), finding an /aj/ raising rate of 84 percent in Canada (equal to the rate of 84 percent found in the acoustic analysis of Boberg, 2008), 80 percent in the U.S. North, and 59 percent in the U.S. West. Although the U.S. West was lower than the other two regions, it was far from lacking raising as would be expected based on Labov et al. (2006). This raised the option of performing the analysis according to participant raising (with their /aj/ raising based on their response to the question and their /aw/ raising assumed based on their region) as opposed to participant region. Both models were tested, but the model based on region was found to be better supported, as judged by the Akaike information criterion (AIC) (Akaike, 1974, 1998; Aho et al., 2014; Kingdom and Prins, 2016), and so the results were primarily interpreted based on region, as initially planned.

Turning to the actual results, the overall region-based pattern found was the opposite of the pattern predicted. Rather than the highest discrimination the U.S. West, followed by the U.S. North, and then Canada, it was Canadians who had the highest accuracy and lowest response times, followed by the U.S. North, and then the U.S. West. Thus, it does not appear to be the case (at least with Canadian Raising) that stronger stereotypes of a dialectal variant are associated with more accurate or easier discrimination. These findings are consistent with Niedzielski (1999), who found that speakers of American English from Detroit listening to the speech of a fellow resident of Detroit were highly influenced in their perception of the speaker's vowels (how raised they were) by whether they were told that the speaker was from Michigan or from Canada. Together, these results suggest that noticing stereotypical dialectal features (at least in relatively similar dialects) can be driven to a large extent by pre-existing knowledge of the speaker's dialect and not necessarily by a heightened sensitivity to those features in the incoming acoustic signal. This conclusion is also supported by two other observations that can be made: the fact that raising plays a prominent role in stereotypes of Canadian English even though it is common in the United States as well (at least for /aj/), and the fact that the *oot and aboot* stereotype is phonetically inaccurate (it portrays raised diphthongs as monophthongs).

Finally, the finding that the model based on region better accounted for the results than the model based on production (both for accuracy and reaction time) suggests that region is important not just for the production differences between the regions, but also for some other factor as well. This factor was explained as being most likely dialect exposure. Canadians presumably have the highest level of exposure to raised variants, while in the United States it is likely that those in the U.S. North are more likely than those in the U.S. West to encounter Canadian Raising in the speech of others, certainly from other speakers in their own dialect region (if raising is more common in the U.S. North than the U.S. West, those who live in the U.S. North will encounter it more often whether or not they have it in their own speech or not) and possibly also from speakers from other regions with high levels of raising, like Canada.

7.4 Online Data Collection

Methodologically, this dissertation relied heavily on web-based remote data collection, with experiments designed and implemented using the jsPsych JavaScript library (de Leeuw, 2015) and participants recruited using the Prolific online subject pool (Palan and Schitter, 2018). There is a fairly extensive literature showing that online data collection is viable and in fact potentially advantageous, due to the ability to recruit larger sample sizes in less time than traditional methods, as well as the ability to test a more diverse sample compared to the traditional method in psycholinguistics and psychology more broadly of relying primarily on undergraduate students as participants (Gosling et al., 2004; Buhrmester et al., 2011; Mason and Suri, 2012; Paolacci and Chandler, 2014; Hauser and Schwarz, 2016; Buhrmester et al., 2018). The first experiment of this dissertation was tested on 32 participants in-lab and an additional 47 participants online, with the participant source being included in the statistical analysis to compare their results. The results were largely equivalent, with the most notable difference being a relatively small accuracy benefit (3 percentage points) for the in-lab participants, which could reflect their linguistics training (many of them were participating for course credit in a linguistics course) rather than a difference in attention or motivation. Afterwards, all data collection was done using online experiments and the results showed no indication of being negatively affected by this source of data, which is to say that there were no major or unexpected drop-offs in performance in later experiments, and later experiments continued to find differences between experimental conditions, with some findings (like the advantage for /aj/ over /aw/) being consistently found across experiments.

This dissertation benefited from online data collection in various specific ways. A total of 623 participants were tested remotely over the course of eight months, which would not have been possible if in-lab testing was continued past Experiment 1. In particular, this dissertation benefited from the ability to easily and quickly re-test experiments with minor modifications to clarify unexpected results (the sub-experiments presented in Chapter 5) and the ability to test non-local dialect groups (the U.S. North and U.S. West in

Chapter 6). These experiences suggest that the field of linguistics would benefit from increased adoption of web-based data collection (including online psycholinguistic experiments, at least when special equipment like an eye-tracker is not needed) because it would increase the accessibility of larger sample sizes (especially for graduate students and other researchers who do not generally have research assistants) and it would increase the accessibility of testing non-local languages and dialects, including under-studied ones.

Appendix A

Experiment 1 stimuli

In Chapter 2, the stimuli from Experiment 1 were visualized as formant trajectories (based on two data points: 20 percent and 80 percent) on the vowel space, with F1 inversely mapped to the y-axis for vowel height and F2 inversely mapped to the x-axis for vowel backness (see Figure 2.6). Here, the same vowels are visualized in a format more resembling a spectrogram, with a smoothed line plotted based on all six data points recorded from each vowel (0, 20, 40, 60, 80, and 100 percent). Note that these are meant to illustrate the stimuli in this experiment rather than to provide a study of the vowel systems of the speakers, and as such they are based on a relatively small number of tokens (only the recordings of the speakers used for stimuli, and not all of the recordings done of the speakers).

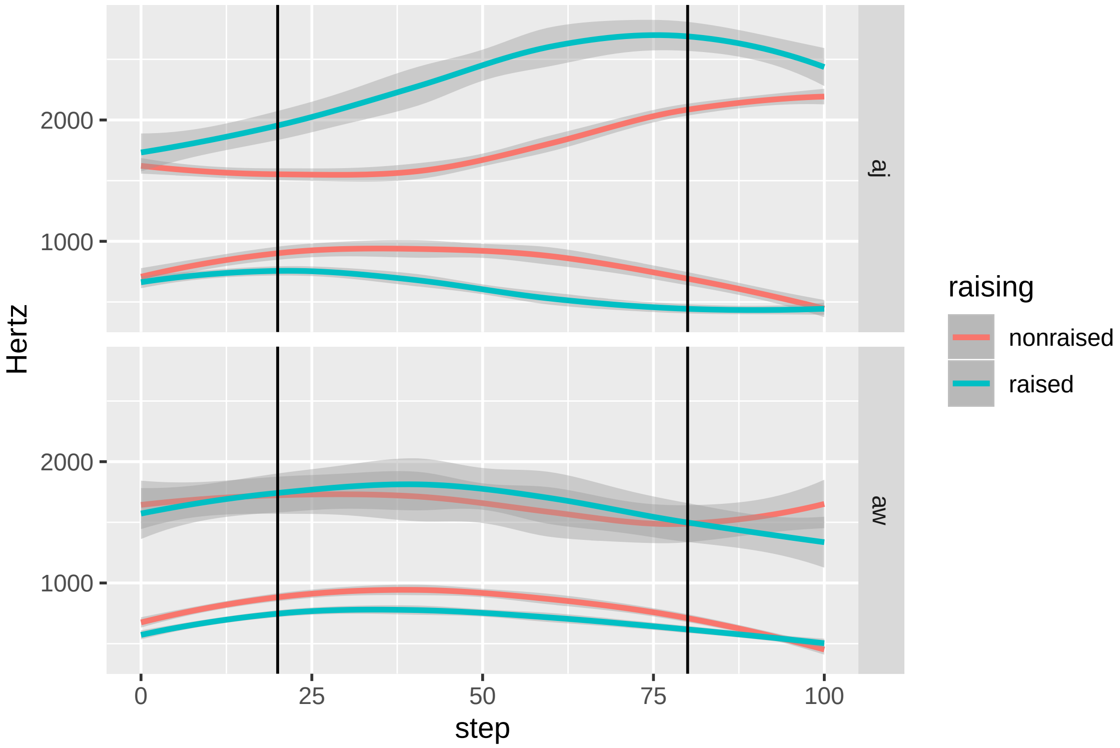

Figure A.1: Smoothed vowel formants (F1 and F2) based on values at 0, 20, 40, 60, 80, and 100 percent from Experiment 1 stimuli (female speaker)

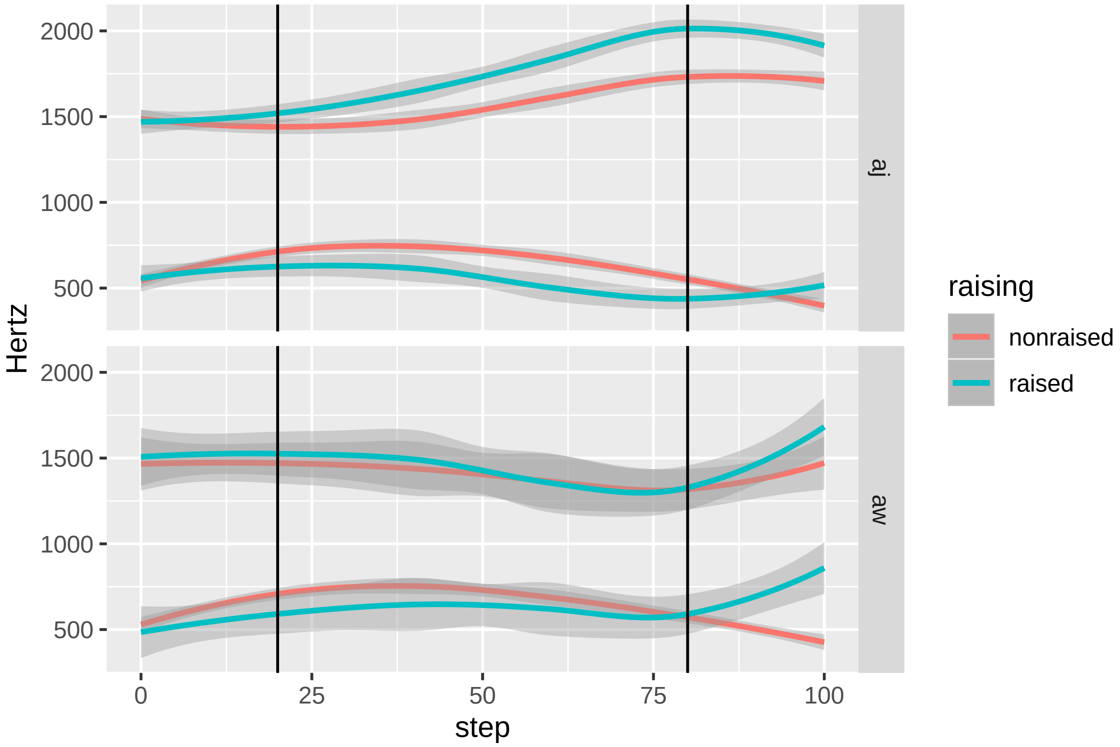

Figure A.2: Smoothed vowel formants (F1 and F2) based on values at 0, 20, 40, 60, 80, and 100 percent from Experiment 1 stimuli (male speaker)

Appendix B

Experiment 2 stimuli

Similarly to the Experiment 1 stimuli, the Experiment 2 stimuli are visualized here in a format more resembling a spectrogram. See Figure 2.7 for the original two-point visualization in vowel space.

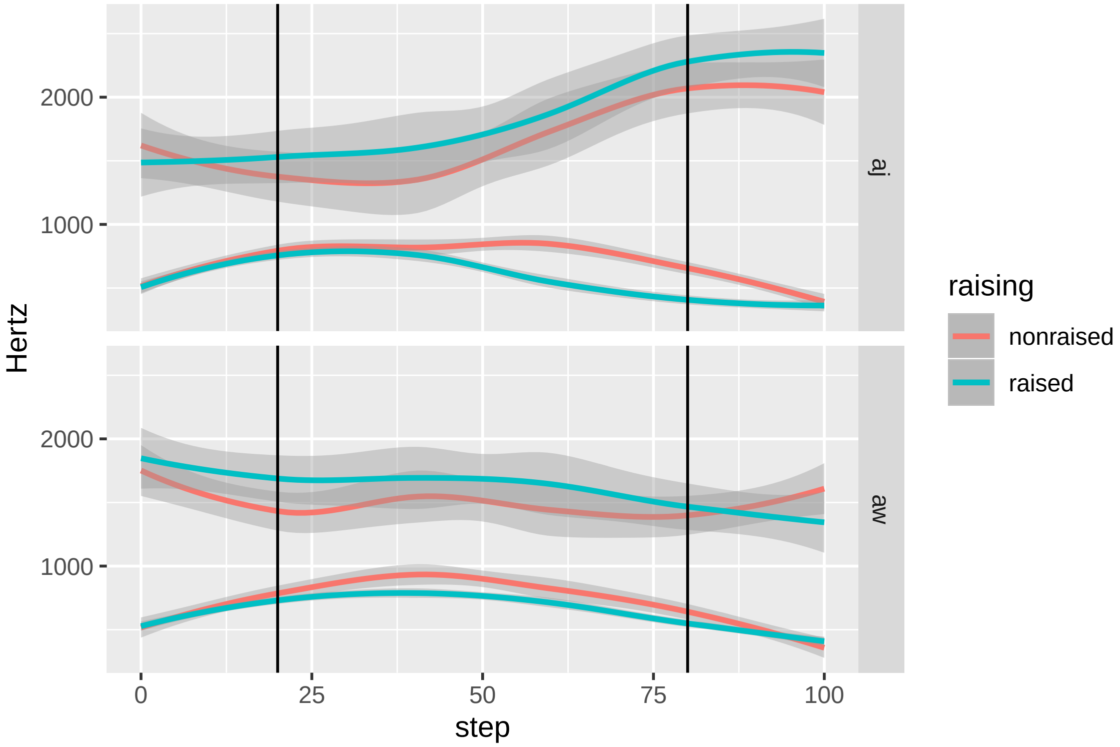

Figure B.1: Smoothed vowel formants (F1 and F2) based on values at 0, 20, 40, 60, 80, and 100 percent from Experiment 2 stimuli (female speaker)

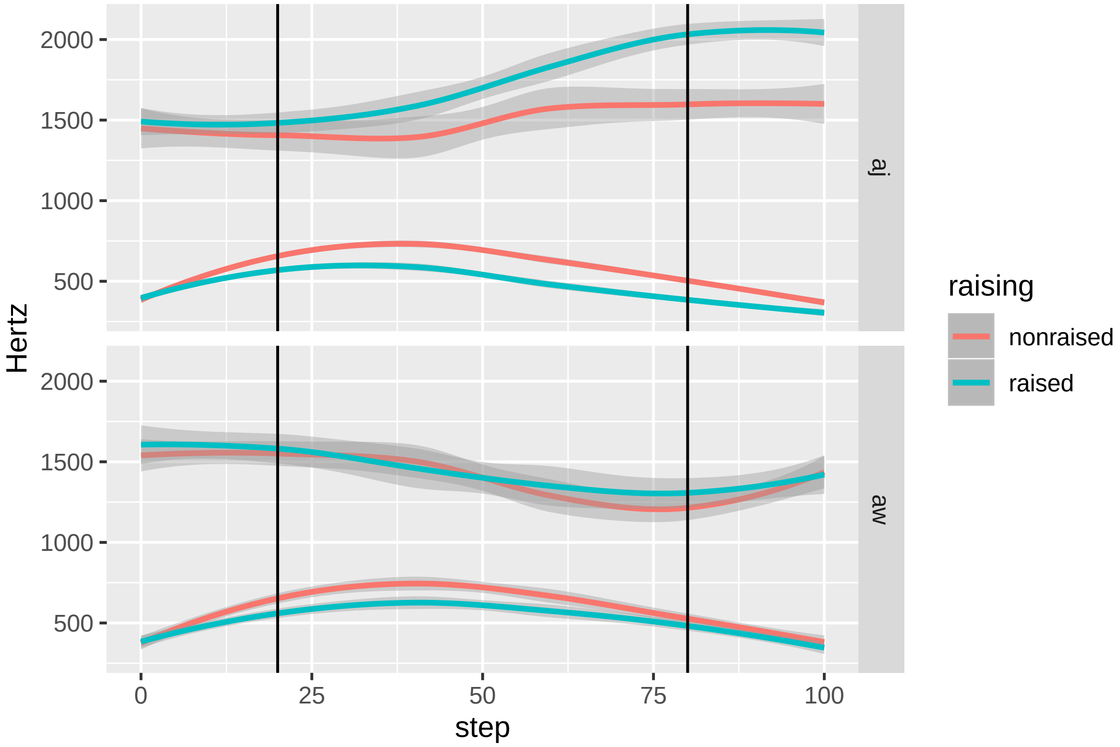

Figure B.2: Smoothed vowel formants (F1 and F2) based on values at 0, 20, 40, 60, 80, and 100 percent from Experiment 2 stimuli (male speaker)

Appendix C

Experiment 5 stimuli

Similarly to the stimuli from Experiments 1 and 2, the Experiment 5 stimuli are visualized here in a format more resembling a spectrogram. See Figure 2.8 for the original two-point visualization in vowel space.

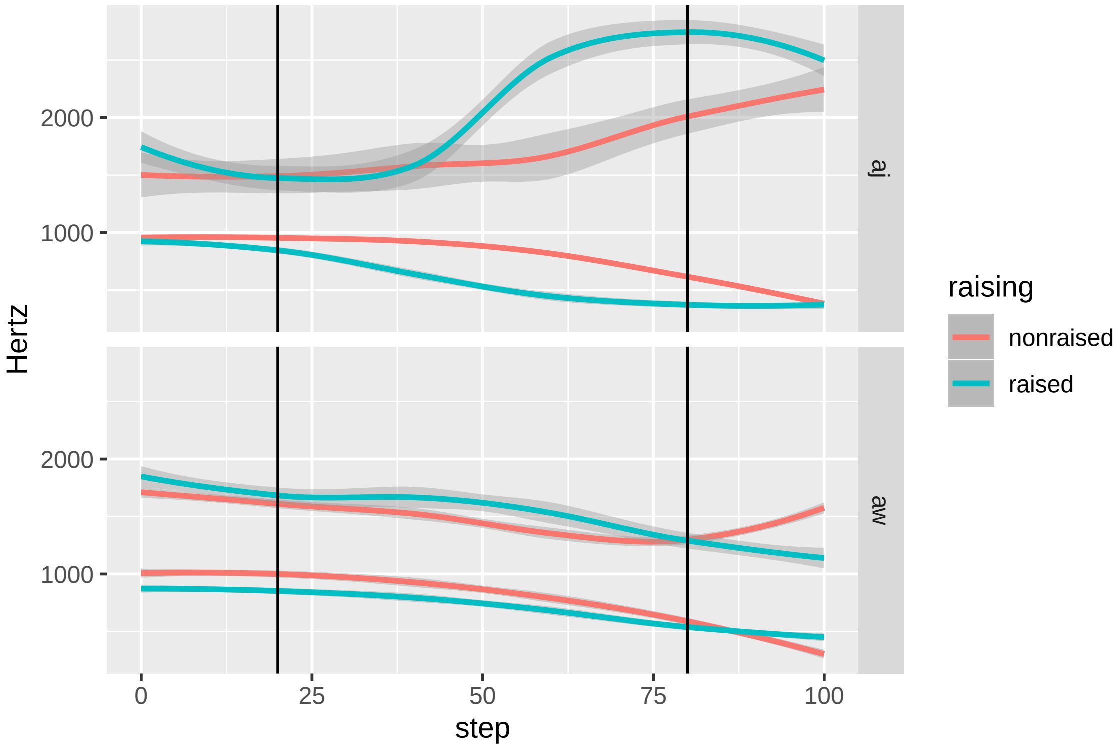

Figure C.1: Smoothed vowel formants (F1 and F2) based on values at 0, 20, 40, 60, 80, and 100 percent from Experiment 5 stimuli (female speaker)

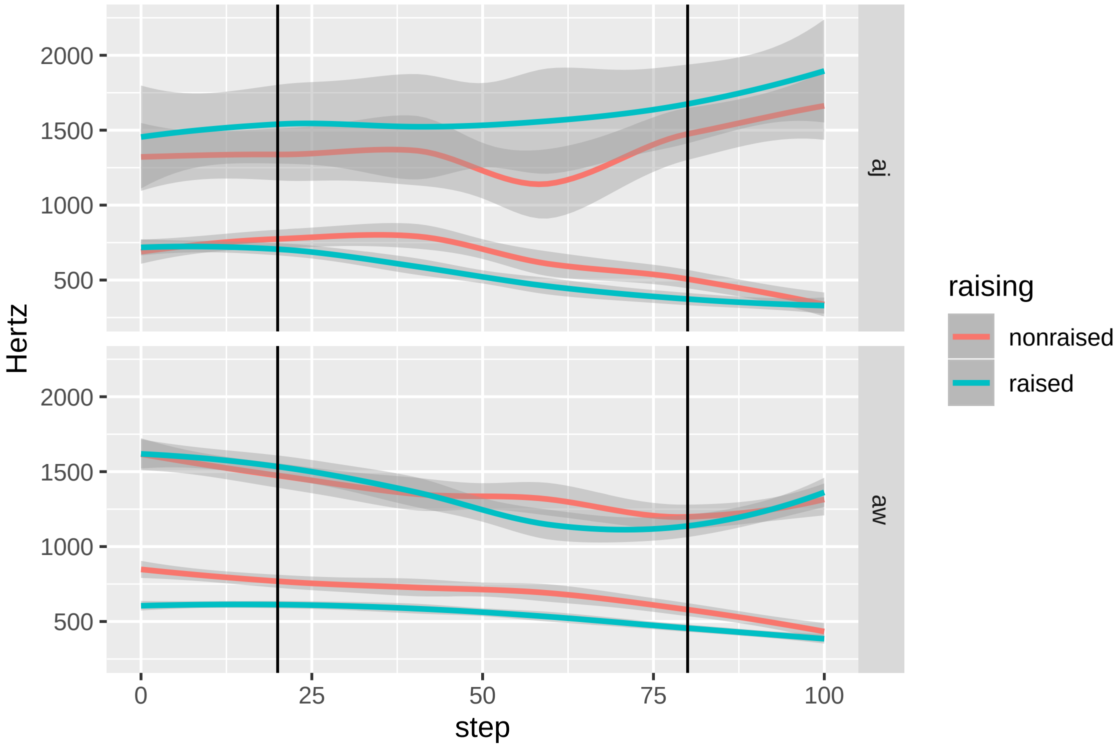

Figure C.2: Smoothed vowel formants (F1 and F2) based on values at 0, 20, 40, 60, 80, and 100 percent from Experiment 5 stimuli (male speaker)

Appendix D

Experiment 5 models

Experiment 5 investigated the discrimination of raised and non-raised diphthongs by Canadians and two groups of Americans (U.S. North, or Great Lakes region, and U.S. West). Two analyses for the results were presented and discussed, with the first one dividing participants up according to region (resulting in three groups: Canada, U.S. North, and U.S. West) and the second dividing participants up according to their raising production (resulting in three groups: raisers of both /aj/ and /aw/, raisers of only /aj/, and non-raisers), which was based on responses to a short dialect survey as well as inferences based on their region. For space reasons, the output for the mixed effects models are presented here. Tables D.1 and D.2 provide the models for accuracy and response time according to participant region, while Tables D.3 and D.4 similarly provide the models for accuracy and response time based on participant raising.

To review the properties of the mixed effects models used, the fixed effects were environment (three levels: isolation, voiceless, and flap), diphthong (two levels: aj and aw), and either participant region or participant raising. The contrast coding used for these categorical variables was simple coding, which provides ANOVA-like main effects rather than simple effects. The reference level was "voiceless" for environment, "aj" for diphthong, "Canada" for participant region, and "both" (i.e., raisers of both /aj/ and /aw/) for participant raising. The approach to random effects was to use the maximal structure justified by the experimental design that does not result in a failure to converge or a singular fit (Barr et al., 2013), but to keep the random effects structure the same between the participant region and participant raising analyses to avoid introducing unnecessary differences for model comparison. The result was by-subjects random intercepts with no random slopes for the accuracy data, and by-subjects random intercepts (with

Table D.1: Mixed effects logistic regression model for Experiment 5 (accuracy) by participant region

term	estimate	std.error	statistic	p.value
(Intercept)	0.650	0.033	19.928	0.000
environmentisolation	0.022	0.039	0.567	0.570
environmentvoiced	0.341	0.040	8.580	0.000
diphthongaw	-0.331	0.032	-10.269	0.000
regionUSNorth	-0.025	0.080	-0.311	0.756
regionUSWest	-0.144	0.080	-1.800	0.072
environmentisolation:diphthongaw	0.095	0.078	1.226	0.220
environmentvoiced:diphthongaw	0.427	0.079	5.371	0.000
environmentisolation:regionUSNorth	0.244	0.096	2.550	0.011
environmentvoiced:regionUSNorth	0.141	0.098	1.433	0.152
environmentisolation:regionUSWest	0.257	0.095	2.699	0.007
environmentvoiced:regionUSWest	0.056	0.097	0.578	0.563
diphthongaw:regionUSNorth	-0.068	0.079	-0.853	0.394
diphthongaw:regionUSWest	0.232	0.079	2.947	0.003
environmentisolation:diphthongaw:regionUSNorth	-0.235	0.191	-1.228	0.219
environmentvoiced:diphthongaw:regionUSNorth	-0.329	0.196	-1.679	0.093
environmentisolation:diphthongaw:regionUSWest	-0.169	0.190	-0.886	0.376
environmentvoiced:diphthongaw:regionUSWest	-0.447	0.195	-2.294	0.022

random slopes for diphthong) for the response time data. No random effect for item was included because each diphthong/environment combination only had one item. The goal of performing a model comparison also necessitated the use of maximum likelihood (ML) rather than restricted maximum likelihood (REML) in the regression analysis.

Table D.2: Mixed effects linear regression model for Experiment 5 (response times) by participant region

term	estimate	std.error	statistic	p.value
(Intercept)	1825.731	14.363	127.113	0.000
environmentisolation	-12.302	16.693	-0.737	0.461
environmentvoiced	3.526	16.693	0.211	0.833
diphthongaw	-13.151	14.283	-0.921	0.359
regionUSNorth	339.960	35.123	9.679	0.000
regionUSWest	444.533	35.300	12.593	0.000
environmentisolation:diphthongaw	30.536	33.387	0.915	0.360
environmentvoiced:diphthongaw	51.200	33.387	1.534	0.125
environmentisolation:regionUSNorth	-20.555	40.822	-0.504	0.615
environmentvoiced:regionUSNorth	-46.320	40.822	-1.135	0.257
environmentisolation:regionUSWest	-52.580	41.027	-1.282	0.200
environmentvoiced:regionUSWest	-68.295	41.027	-1.665	0.096
diphthongaw:regionUSNorth	6.215	34.928	0.178	0.859
diphthongaw:regionUSWest	-7.543	35.104	-0.215	0.830
environmentisolation:diphthongaw:regionUSNorth	-43.610	81.643	-0.534	0.593
environmentvoiced:diphthongaw:regionUSNorth	-54.938	81.643	-0.673	0.501
environmentisolation:diphthongaw:regionUSWest	-39.009	82.055	-0.475	0.635
environmentvoiced:diphthongaw:regionUSWest	-138.894	82.055	-1.693	0.091

Table D.3: Mixed effects logistic regression model for Experiment 5 (accuracy) by participant raising

term	estimate	std.error	statistic	p.value
(Intercept)	0.663	0.035	19.205	0.000
environmentisolation	-0.009	0.041	-0.224	0.823
environmentvoiced	0.328	0.042	7.783	0.000
diphthongaw	-0.310	0.034	-9.104	0.000
raiseronlyaj	-0.127	0.077	-1.651	0.099
raiserneither	-0.063	0.093	-0.684	0.494
environmentisolation:diphthongaw	0.136	0.082	1.654	0.098
environmentvoiced:diphthongaw	0.442	0.084	5.243	0.000
environmentisolation:raiseronlyaj	0.221	0.092	2.401	0.016
environmentvoiced:raiseronlyaj	0.056	0.094	0.595	0.552
environmentisolation:raiserneither	0.023	0.110	0.207	0.836
environmentvoiced:raiserneither	-0.044	0.113	-0.388	0.698
diphthongaw:raiseronlyaj	0.070	0.076	0.918	0.359
diphthongaw:raiserneither	0.279	0.092	3.047	0.002
environmentisolation:diphthongaw:raiseronlyaj	-0.193	0.184	-1.049	0.294
environmentvoiced:diphthongaw:raiseronlyaj	-0.269	0.188	-1.427	0.153
environmentisolation:diphthongaw:raiserneither	0.118	0.220	0.535	0.593
environmentvoiced:diphthongaw:raiserneither	-0.240	0.226	-1.059	0.290

Table D.4: Mixed effects linear regression model for Experiment 5 (response times) by participant raising

term	estimate	std.error	statistic	p.value
(Intercept)	1801.384	17.451	103.223	0.000
environmentisolation	-0.219	17.626	-0.012	0.990
environmentvoiced	12.457	17.626	0.707	0.480
diphthongaw	-17.815	15.026	-1.186	0.238
raiseronlyaj	371.701	38.937	9.546	0.000
raiserneither	313.172	46.768	6.696	0.000
environmentisolation:diphthongaw	22.448	35.253	0.637	0.524
environmentvoiced:diphthongaw	53.568	35.253	1.520	0.129
environmentisolation:raiseronlyaj	-49.881	39.328	-1.268	0.205
environmentvoiced:raiseronlyaj	-65.242	39.328	-1.659	0.097
environmentisolation:raiserneither	36.970	47.237	0.783	0.434
environmentvoiced:raiserneither	-13.114	47.237	-0.278	0.781
diphthongaw:raiseronlyaj	5.540	33.525	0.165	0.869
diphthongaw:raiserneither	-34.640	40.267	-0.860	0.391
environmentisolation:diphthongaw:raiseronlyaj	-26.647	78.655	-0.339	0.735
environmentvoiced:diphthongaw:raiseronlyaj	-81.194	78.655	-1.032	0.302
environmentisolation:diphthongaw:raiserneither	-110.918	94.474	-1.174	0.240
environmentvoiced:diphthongaw:raiserneither	-96.782	94.474	-1.024	0.306

Bibliography

Abramson, A. S. (1977). Noncategorical perception of tone categories in Thai. *The Journal of the Acoustical Society of America*, 61(S1):S66–S66.

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3):631–636.

Ahrend, E. R. (1934). Ontario Speech. *American Speech*, 9(2):136–139.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In Parzen, E., Tanabe, K., and Kitagawa, G., editors, *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 199–213. Springer New York, New York, NY.

Allbritten, R. M. (2011). *Sounding Southern : phonetic features and dialect perceptions*. PhD, Georgetown University, Washington, D.C.

Allen, H. B. (1989). Canadian Raising in the Upper Midwest. *American*, 64(1):74–75.

Aperlinski, G. and Schwartz, G. (2015). Release bursts vs. formant transitions in Polish stop place perception. In *ICPhS*.

Austin, P. K. (1988). Phonological voicing contrasts in Australian Aboriginal languages. *La Trobe Working Papers in Linguistics*, 1:17–42.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3):445–459.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3).

Barrios, S. L., Namyst, A. M., Lau, E. F., Feldman, N. H., and Idsardi, W. J. (2016). Establishing New Mappings between Familiar Phones: Neural and Behavioral Evidence for Early Automatic Processing of Nonnative Contrasts. *Frontiers in Psychology*, 7.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Berkson, K. and Herd, W. (2017). Incipient /aɪ/-raising in Baton Rouge.

Best, C. C. and McRoberts, G. W. (2003). Infant Perception of Non-Native Consonant Contrasts that Adults Assimilate in Different Ways. *Language and speech*, 46(Pt 2-3):183–216.

Best, C. T., McRoberts, G. W., and Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2):775–794.

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology. Human Perception and Performance*, 14(3):345–360.

Best, C. T., Traill, A., Carter, A., Harrison, K. D., and Faber, A. (2003). !xóõ click perception by english, isizulu, and sesotho listeners. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 853–856.

Boberg, C. (2008). Regional Phonetic Differentiation in Standard Canadian English. *Journal of English Linguistics*, 36(2):129–154.

Boersma, P. and Weenick, D. (2017). Praat: doing phonetics by computer.

Boomershine, A., Currie, K., Hume, E., and Johnson, K. (2008). The Impact of Allophony versus Contrast on Speech Perception.

Bornstein, M. H. and Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: some implications for categorical perception and levels of information processing. *Psychological Research*, 46(3):207–222.

Braver, A. (2013). *Degrees of Incompleteness in Neutralization: Paradigm Uniformity in a Phonetics with Weighted Constraints.* PhD, Rutgers, The State University of New Jersey.

Braver, A. (2014). Imperceptible incomplete neutralization: Production, non-identifiability, and non-discriminability in American English flapping. *Lingua*, 152:24–44.

Bridgeman, E. and Snowling, M. (1988). The perception of phoneme sequence: A comparison of dyspraxic and normal children. *International Journal of Language & Communication Disorders*, 23(3):245–252.

Britain, D. (1997). Dialect Contact and Phonological Reallocation: "Canadian Raising" in the English Fens. *Language in Society*, 26(1):15–46.

Broersma, M. (2010). Perception of final fricative voicing: Native and nonnative listeners' use of vowel duration. *The Journal of the Acoustical Society of America*, 127(3):1636–1644.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Buhrmester, M. D., Kwang, T. N., and Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on psychological science : a journal of the Association for Psychological Science*, 6(1):3–5.

Buhrmester, M. D., Talaifar, S., and Gosling, S. D. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, 13(2):149–154.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304.

Burns, E. M. and Ward, W. (1974). Categorical Perception of Musical Intervals. *Acoustical Society of America Journal*, 55:456.

Cardoso, A. B. (2015). *Dialectology, phonology, diachrony: Liverpool English realisations of PRICE and MOUTH.* PhD, University of Edinburgh, Edinburgh.

Case, J., Seyfarth, S., and Levi, S. V. (2018a). Does Implicit Voice Learning Improve Spoken Language Processing? Implications for Clinical Practice. *Journal of Speech, Language, and Hearing Research : JSLHR*, 61(5):1251–1260.

Case, J., Seyfarth, S., and Levi, S. V. (2018b). Short-term implicit voice-learning leads to a Familiar Talker Advantage: The role of encoding specificity. *The Journal of the Acoustical Society of America*, 144(6):EL497–EL502.

Celata, C. (2008). Partial phonological contrasts in native and non-native speech perception.

Chambers, J. (1973). Canadian Raising. *Canadian Journal of Linguistics*, 18(2):113–135.

Chambers, J. (2006a). Canadian Raising Retrospect and Prospect. *The Canadian Journal of Linguistics / La revue canadienne de linguistique*, 51:105–118.

Chambers, J. K. (1994). An Introduction to Dialect Topography. *English World-Wide*, 15(1):35–53.

Chambers, J. K. (2006b). Geolinguistic Patterns in a Vast Speech Community. *Linguistica Atlantica*, 27-28:27–36.

Chambers, J. K. and Hardwick, M. F. (1986). Comparative Sociolinguistics of a Sound Change in Canadian English. *English World-Wide*, 7(1):23–46.

Chambers, J. K. and Heisler, T. (1999). Dialect Topography of Québec City English. *Canadian Journal of Linguistics; Toronto, etc.*, 44(1):23.

Chan, S. W., Chuang, C., and Wang, W. S. (1975). Cross-linguistic study of categorical perception for lexical tone. *The Journal of the Acoustical Society of America*, 58(S1):S119–S119.

Chen, M. (1970). Vowel Length Variation as a Function of the Voicing of the Consonant Environment. *Phonetica*, 22(3):129–159.

Cho, H. (2016). Variation in vowel duration depending on voicing in American, British, and New Zealand English. *Phonetics and Speech Sciences*, 8:11–20.

Cho, T., Jun, S.-A., and Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, 30(2):193–228.

Cho, T., Whalen, D., and Docherty, G. (2019). Voice onset time and beyond: Exploring laryngeal contrast in 19 languages. *Journal of Phonetics*, 72:52–65.

Corretge, R. (2012). Praat Vocal Toolkit.

Dailey-O'Cain, J. (1997). Canadian raising in a midwestern U.S. city. *Language Variation and Change*, 9:107–120.

Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America*, 120(1):407–415.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.

de Leeuw, J. R. (2015). jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1):1–12.

Dresher, B. E. (2011). *The Contrastive Hierarchy in Phonology*. Cambridge University Press. Google-Books-ID: aFEvcgAACAAJ.

Eddington, D. and Taylor, M. (2009). T-Glottalization in American English. *American Speech*, 84:298–314.

Etcoff, N. L. and Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44(3):227–240.

Fischer-Jorgensen, E. (1972). Tape Cutting Experiments with Stop Consonants. *The Journal of the Acoustical Society of America*, 52(1A):132–132.

Fiske, H. E. (1997). Categorical Perception of Musical Patterns: How Different Is ”Different”. *Bulletin of the Council for Research in Music Education*, (133):20–24.

Fox, R. A. and Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America*, 126(5):2603–2618.

Francis, A. L., Ciocca, V., and Ng, B. K. C. (2003). On the (non)categorical perception of lexical tones. *Perception & Psychophysics*, 65(7):1029–1044.

Fruehwald, J. (2008). The Spread of Raising: Opacity, Lexicalization, and Diffusion. *University of Pennsylvania Working Papers in Linguistics*, 14(2).

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). The Identification and Discrimination of Synthetic Vowels. *Language and Speech*, 5(4):171–189.

Fugate, J. M. B. (2013). Categorical Perception for Emotional Faces. *Emotion review : journal of the International Society for Research on Emotion*, 5(1):84–89.

Fujisaki, H. and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute (Faculty of Engineering, University of Tokyo)*, 29:207–214.

Fullerton, G. (2019). Canadian Raising: Contrast Neutralization in Meaford, ON and Vancouver, BC. Vancouver, BC.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, 6(1):110–125.

Glosser, G., Kohn, S., Friedman, R., Sands, L., and Grugan, P. (1997). Repetition of Single Words and Nonwords in Alzheimer's Disease. *Cortex*, 33(4):653–666.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 22(5):1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105(2):251–279.

Goldinger, S. D., Kleider, H. M., and Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition*, 27(2):328–338.

Goldstone, R. L. and Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78.

Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *The American Psychologist*, 59(2):93–104.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia*, 9(3):317–323.

Hall, D. C. and Hall, K. C. (2016). Marginal contrasts and the Contrastivist Hypothesis. *Glossa: a journal of general linguistics*, 1(1):1–23.

Hall, E. (2015). Static and Dynamic Analyses of Canadian Raising in Toronto and Vancouver.

Hall, K. C. (2005). Defining Phonological Rules over Lexical Neighbourhoods: Evidence from Canadian Raising. In *Proceedings of the 24th West Coast Conference on Formal Linguistics*, pages 191–199, Simon Fraser University (Vancouver). Cascadilla Proceedings Project.

Hall, K. C. (2009). *A Probabilistic Model of Phonological Relationships from Contrast to Allophony.* PhD, The Ohio State University, Columbus, Ohio.

Hall, K. C. (2012). Phonological Relationships: A Probabilistic Model. *McGill Working Papers in Linguistics*, 22(1).

Hall, K. C. (2013). A typology of intermediate phonological relationships. *The Linguistic Review*, 30(2):215–275.

Hamre, C. (2019). Interspeaker and intraspeaker variation in Buffalo Canadian Raising. 2nd Annual Buffalo-Toronto Workshop on Linguistic Perspectives on Variation Within and Across Languages.

Han, M. S. and Weitzman, R. S. (1970). Acoustic Features of Korean /P, T, K/, /p, t, k/ and /ph, th, kh/. *Phonetica*, 22(2):112–128.

Harris, K. S. (1958). Cues for the Discrimination of American English Fricatives in Spoken Syllables. *Language and Speech*, 1(1):1–7.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). Effect of Third-Formant Transitions on the Perception of the Voiced Stop Consonants. *The Journal of the Acoustical Society of America*, 30(2):122–126.

Hauser, D. J. and Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1):400–407.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *The Journal of the Acoustical Society of America*, 108(2):710–722.

Hualde, J. I. (2004). Quasi-Phonemic Contrasts in Spanish. In *Proceedings of the 23rd West Coast Conference on Formal Linguistics*, pages 374–398, University of California, Davis. Cascadilla Press.

Huang, T. (2010). Why are vowels and tones perceived less categorically than stop consonants? – Categorical perception and speech sound types revisited.

Hume, E. and Johnson, K. (2003). The Impact of Partial Phonological Contrast on Speech Perception. In *Proceedings of the 15th International Congress of Phonetic Sciences*.

Johnson, K. and Babel, M. (2010). On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers. *Journal of Phonetics*, 38(1):127–136.

Joos, M. (1942). A Phonological Dilemma in Canadian English. *Language*, 18(2):141–144.

Kingdom, F. A. A. and Prins, N. (2016). Chapter 9 - Model Comparisons**This chapter was primarily written by Nicolaas Prins. In Kingdom, F. A. A. and Prins, N., editors, *Psychophysics (Second Edition)*, pages 247–307. Academic Press, San Diego.

Kishon-Rabin, L., Dayan, M., and Michaeli, O. (2011). Effect of Second-Formant Transitions on the Perception of Hebrew Voiced Stop Consonants. *Journal of Basic and Clinical Physiology and Pharmacology*, 14(2):151–164.

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, 23(6):1681–1712.

Kröger, B. J., Birkholz, P., Kannampuzha, J., and Neuschaefer-Rube, C. (2011). Categorical Perception of Consonants and Vowels: Evidence from a Neurophonetic Model of Speech Production and Perception. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, Lecture Notes in Computer Science, pages 354–361. Springer, Berlin, Heidelberg.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). *lmerTest: Tests in linear mixed effects models*.

Labov, W., Ash, S., and Boberg, C. (2006). *Atlas of North American English*. Mouton de Gruyter, Berlin.

Ladd, D. R. (2006). ”Distinctive phones” in surface representation. *Laboratory Phonology*, 8:3–26.

Lago, S., Scharinger, M., Kronrod, Y., and Idsardi, W. J. (2015). Categorical effects in fricative perception are reflected in cortical source information. *Brain and language*, 143:52–58.

Lahiri, A. and Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, 38(3):245–294.

Lee, S. and Katz, J. (2016). Perceptual integration of acoustic cues to laryngeal contrasts in Korean fricatives. *The Journal of the Acoustical Society of America*, 139(2):605–611.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 68(8):1–13.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Lindblom, B. E. and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America*, 42(4):830–843.

Lisker, L. and Abramson, A. (1970). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the sixth International Congress of Phonetic Sciences*.

Luce, P. A. and Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America*, 78(6):1949–1957.

Luce, P. A. and Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, 26(4):708–715.

MacKain, K. S., Best, C. T., and Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, 2(4):369–390.

Macmillan, N. A. and Creelman, C. D. (2004). *Detection Theory: A User's Guide*. Psychology Press, Mahwah, N.J, 2 edition edition.

Magnuson, T. J. (1998). *What /r/ Sounds Like in Kansai Japanese: A Phonetic Investigation of Liquid Variation in Unscripted Discourse*. Bachelor's Thesis, University of British Columbia, Vancouver, BC.

Mann, V. A. and Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, 28(3):213–228.

Mason, W. and Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23.

Mathôt, S., Schreij, D., and Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2):314–324.

McGuire, G. L. (2007). *Phonetic category learning*. PhD, The Ohio State University, The Ohio State University.

McQueen, J. M., Cutler, A., and Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30(6):1113–1126.

Menon, K., Rao, P., and Thosar, R. (1974). Formant Transitions and Stop Consonant Perception in Syllables. *Language and Speech*, 17(1):27–46.

Meredith, M. and Maye, J. (2009). Perception of phonemic and allophonic contrasts in English- and Spanish-learning infants. *The Journal of the Acoustical Society of America*, 125(4):2766–2766.

Mielke, J., Armstrong, M., and Hume, E. (2008). Looking through opacity. *Theoretical Linguistics*, 29(1-2):123–139.

Monahan, P. J. (2009). *On The Way To Linguistic Representation: Neuromagnetic Evidence of Early Auditory Abstraction in the Perception of Speech and Pitch.* PhD, University of Maryland, College Park, College Park, Maryland.

Munson, W. A. and Gardner, M. B. (1950). Standardizing Auditory Tests. *The Journal of the Acoustical Society of America*, 22(5):675–675.

Murphy, P., Monahan, P., and Grant, M. (2016). Affrication Patterns and Perceptual Tendencies in Canadian and European French. *Canadian Linguistic Association 2016 Conference Proceedings.*

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, 85(5):2088–2113.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, 18(1):62–85.

Palan, S. and Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Paolacci, G. and Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science.*

Phillips, C. (2001). Levels of representation in the electrophysiology of speech perception. *Cognitive Science*, 25(5):711–731.

Pi, C.-Y. T. (2000). Canadians Telling Time: A Study in Dialect Topography. *Toronto Working Papers in Linguistics*, 18:80–102.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & psychophysics*, 13(2):253–260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & cognition*, 3(1).

Pitt, M. A. and Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology. Human Perception and Performance*, 19(4):699–725.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163.

Raphael, L. J. (1972). Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English. *The Journal of the Acoustical Society of America*, 56(4B):1296–1303.

Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychophysics*, 30(3):217–227.

Roberson, D., Pak, H., and Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition*, 107(2):752–762.

Roberts, J. (2006). As Old Becomes New: Glottalization in Vermont. *American Speech*, 81(3):227–249.

Roberts, J. (2007). Vermont lowering? Raising some questions about /ai/ and /au/ south of the Canadian border. *Language Variation and Change*, 19:181–197.

Rosenfelder, I. (2007). Canadian Raising in Victoria, B.C.: An Acoustic Analysis. *AAA: Arbeiten aus Anglistik und Amerikanistik*, 32(2):257–284.

Sadlier-Brown, E. (2012). Homogeneity and autonomy of Canadian Raising. *World Englishes*, 31(4):534–548.

Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology. General*, 110(4):474–494.

Schertz, J., Kang, Y., and Han, S. (2019). Sources of variability in phonetic perception: The joint influence of listener and talker characteristics on perception of the Korean stop contrast. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 10(1):13.

Seyfarth, S. and Garellek, M. (2015). Coda glottalization in American English. In *ICPhS*.

Smits, R., ten Bosch, L., and Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. *The Journal of the Acoustical Society of America*, 100(6):3852–3864.

Souza, P., Gehani, N., Wright, R., and McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, 24(8):689–700.

Sproat, R. and Fujimura, O. (1993). Allophonic Variation in English /l/ and Its Implications for Phonetic Implementation. *Journal of Phonetics*, 21:291–311.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43(1):155–167.

Statistics Canada (2015). Table 24-10-0040-01 Travellers to Canada from the United States by state of origin, top 15 states of origin.

Stevenson, S. and Zamuner, T. (2017). Gradient phonological relationships: Evidence from vowels in French. *Glossa: a journal of general linguistics*, 2(1):58.

Stricker, A., Berkson, K., and Davis, S. (2016). Canadian raising in Fort Wayne, Indiana. Salt Lake City, UT.

Studdert-Kennedy, M. (1976). Speech Perception. In Lass, N. J., editor, *Contemporary Issues in Experimental Phonetics*, pages 243–293. Academic Press.

Sundara, M., Polka, L., and Genesee, F. (2006). Language-experience facilitates discrimination of /d-th/ in monolingual and bilingual acquisition of English. *Cognition*, 100(2):369–388.

Swan, J. T. (2016). *Language ideologies, border effects, and dialectal variation: Evidence from /æ/, /aʊ/, and /ai/ in Seattle, WA and Vancouver, BC.* PhD, University of Chicago, Chicago.

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P., and Best, C. (1994). Discrimination of english /r-l/ and /w-y/ by japanese infants at 6-12 months: Language-specific developmental changes in speech perception abilities. In *Third International Conference on Spoken Language Processing*.

U.S. Census Bureau (2019). Nst-est2018-01: Table 1. annual estimates of the resident population for the united states, regions, states, and puerto rico: April 1, 2010 to july 1, 2018.

Vaden, K., Halpin, H., and Hickok, G. (2009). Irvine Phonotactic Online Dictionary, Version 2.0.

Vance, T. (1987). Canadian Raising in Some Dialects of the Northern United States. *American Speech*, 62:3.

Vaux, B. (2000). Flapping in English. Chicago.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science (New York, N.Y.)*, 167(3917):392–393.

Werker, J. F., Gilbert, J. H., Humphrey, K., and Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, 52(1):349–355.

Werker, J. F. and Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37(1):35–44.

Whalen, D. H., Best, C. T., and Irwin, J. R. (1997). Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics*, 25(4):501–528.

Williams, D. R. (1987). Judgments of coarticulated vowels are based on dynamic information. *The Journal of the Acoustical Society of America*, 81(S1):S17–S17.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, 104(19):7780–7785.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., and Perrett, D. I. (1997). Facial expression megamix: tests of dimensional and category accounts of emotion recognition. *Cognition*, 63(3):271–313.

Zhang, X. and Samuel, A. G. (2015). The Activation of Embedded Words in Spoken Word Recognition. *Journal of memory and language*, 79-80:53–75.