UNIVERSITY OF CALIFORNIA

Los Angeles

Categorical and non-categorical perception of marginal phonemes

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Linguistics

by

Zhenglong Zhou

2024

ABSTRACT OF THE DISSERTATION

Categorical and non-categorical perception of marginal phonemes

by

Zhenglong Zhou
Doctor of Philosophy in Linguistics
University of California, Los Angeles, 2024
Professor Patricia Keating, Chair

Marginal phonemes and contrasts occupy a complex position in linguistic theory, as traditional theories of phonemehood do not account for marginality. However, contemporary linguistics has found that phonemic contrast strength is not fixed in childhood but rather continues to change, thus implicating the lexicon, the set of words a speaker knows, as a factor in the behavior of phonemic contrasts.

This dissertation takes this link between category strength and the lexicon and treats it as an empirical question. I identify token frequency and type informativity, measures of frequency and predictability within a lexicon, as potential predictors of individual behavior. I then justify and present an experimental procedure for an eye tracking, two-alternative forced choice, categorization study on three phonetic continua — [a͡ɪ]-[ʌ͡i], a marginal contrast; [a͡ɪ]-[ɔ͡ɪ], a classic phonemic contrast; and [ʌ͡i]-[ɔ͡ɪ], a mixed case — in Canadian English, using the visual world paradigm. I discuss decisions that were made in the design of the experiment, including how individual lexicons were probed and why multiple continua were examined.

Analyzing the resultant eye tracking data both graphically and by GAMM model comparison, I find that behavior was not interpretably predicted by my selected predictors, though their contribution to bias, a normalized preference measure, was statistically significant. I report on the behavioral patterns that were found in the categorization data and show that participants with differential behavior did not have statistically significant differences in either frequency or informativity in nearly all cases.

My findings come as a surprise, as predicting that variation in the lexicon (operationalized as frequency and informativity) should influence linguistic behavior is both obvious and supported by the literature. I thus present my thoughts on why these predictors were not significant ones as well as my suspicion that the process of calculating these lexical statistics was poisoned by the likely incorrect assumption that a marginal phoneme can be treated as if it were a strong phoneme for the purposes of calculation. I close with suggestions for future work that could advance understanding of this issue, including potential test cases and the need for alternative operationalizations of frequency and predictability for marginal phonemes.

The dissertation of Zhenglong Zhou is approved.

Kie Ross Zuraw

Megha Sundara

Bruce P. Hayes

Patricia Keating, Committee Chair

University of California, Los Angeles

2024

For my family,

those who I love

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Looking back at my time at the UCLA Linguistics department, I cannot feel anything but extreme gratitude for the advice, encouragement, assistance, and friendship of the wonderful people it has been my sincere pleasure to have shared a portion of my life with. And how! These years have been some of the best, most formulative years of my life.

As is tradition, I have to first thank my committee. I am incredibly grateful to Pat Keating, my advisor and chair, for her wisdom and measured perspective. Pat, not only was it a pleasure to hear your insights into my work and enjoy your support as I sought to develop my research interests and then my professional ones, I treasure the time we spent as department chair and GLC president. The deftness and skill of your approach to so many matters, personal, professional, political, intellectual, and your commitment to your students — thanks for being an inspiration.

Of course, I am no less grateful to my other committee members. Kie, the advice you gave me when I started this dissertation to 1) treat it as just another (big) project and 2) keep copious notes on my thought process has been some of the most timely counsel I have ever received. Thank you; I would not have been able to finish this dissertation otherwise. Bruce, though you have given me so much to consider over the years, intellectually, I am most grateful to you for pushing me to remain grounded while still encouraging my dreams: "concretize" is one of my favorite words because of you. Megha, I remember when you strategized with me in my second year about how to position my research in a cohesive narrative. Though that plan didn't work out (because of me), I was touched by your sincerity, your savvy, and your desire to see me succeed. Thank you for the intensity and clarity of purpose you have shown me.

Of course, of course!, I must thank Ashley Farris-Trimble and Danica Reid. The experimental portion of my dissertation took place in Ashley's Phonological Processing Lab, and, Ashley, I am forever grateful to you for your experience, for letting me use your facilities and resources, and for your help in setting up an international IRB. But more than anything, thanks for your empathy. Danica, thanks for your contributions to the experiment. Thanks for writing the logic for the training and testing blocks. Not only was your experience with ExperimentBuilder a godsend early on, it was also a lifesaver later, when we had to deal with the offset issue. Also, thanks for being the voice of the study — I know it was rough. A sincere thanks, too, to the PPL research assistants who were the arms and legs of this study, running participants and checking wordlists and code: Allyson Ugalde, Kylie Brajcich, Sahibnoor Dhami, and Chelsea Riebesehl McGarvie.

I would like to also thank my past mentors from Swarthmore, Donna Jo Napoli, Nathan Sanders, and Byron Ahn. I would never have come to this degree, this program, this field, without your generous, generous attention and support. It is impossible for me to overstate how your influence has altered my direction — I can literally not imagine where I would be in life if not for you. Thank you for modeling excellence in mentorship for me.

Thanks, now, to the fellow travellers who have accompanied me along this journey from BA to PhD. First, to Rachel Vogel and Canaan Breiss. You know. To my cohortmates, Hironori Katsuda and Christine Prechtel, Maddy Booth and Minqi Liu, and Andy Xu. I will always look back fondly on the many good times we spent together, in class, on the bus, in Australia; eating, drinking, kiki-ing, kvetching; discussing the nature of science, science fiction, justice, virtue — and on the hard times we spent together, as well. To Meng Yang and Deborah Wong, thanks for being there and for showing me LA. To quote Meng, "because of you, I was much less productive, but much happier." A special thanks to Jeremy Steffman, a most kind soul, for showing me how to analyze eye tracking data when I had never done such a thing before. To Jinyoung Jo, Jennifer Kuo, and Angelica Pan, thanks for coming to all the parties. To those members of the P-lab and the department who I whiled away so many hours with, making scintillating conversation: Jesse Zymet, Marju Kaps, Travis Major, Connor Mayer, Noah Elkins, Jake Aziz, Jahnavi Narkar, and Jian-Leat Siah. The past few years would have been awfully dry and boring without you. Thank you, for engaging.

Alas, it would be impossible to thank everyone who has influenced me on the way here; my old friends from UHS and Swarthmore, new(er) friends from LA and SF... there are too many of you, and this is not the place for long, winding stories about cause and effect, besides. So I must regretfully limit myself now to thanking those with specific, yet far-reaching, impact on my life.

Thanks to DeAnna McDonald and Christopher Nelson. The lessons I learned in your classrooms have never left me. And thanks to Ruth Chau, who, at a crossroads in my life, was the first person to *really* ask me what I want, and to Lauren Northrup, who, at a different inflection point many years later, gave me the tools to find a path north.

Finally, I thank my family for their support over these past several years, always looking out for my health and well-being, my career prospects, and my social life. Most specifically, thanks to Brandon Sixto, o lodestar, o light in the night — thanks for everything.

VITA

2019	M.A. Linguistics, University of California, Los Angeles
2019	NSF Graduate Research Fellowship Program Honorable Mention
2016	B.A. in Linguistics, Swarthmore College

PUBLICATIONS AND PRESENTATIONS

Ahn, Byron, Z.L. Zhou, Emily Gasser & Donna Jo Napoli. (To appear). Perception of the prosodic features of newscaster speech. In *A festschrift for Jack Hoeksema*.

Minkova, Donka & Z.L. Zhou. (2023). Nominal compounds in OE meter and prosody. Journal of Germanic Linguistics 35(1): 69–95. doi: 10.1017/S1470542722000083.

Minkova, Donka & Z.L. Zhou. (2022). Early metrical and lexicographical evidence for functional stress-shifts. English Language & Linguistics 26(3): 533–558. doi: 10.1017/S1360674322000144.

Sundara, Megha, Z.L. Zhou, Canaan Breiss, Hironori Katsuda, & Jeremy Steffman. (2022). Infants' developing sensitivity to native language phonotactics: A meta-analysis. Cognition 221: 104993. doi: 10.1016/j.cognition.2021.104993.

Zhou, Z.L. & Canaan Breiss. (2021). Towards substantively biased typology: Effects of environment on P-map biases. Presented at Linguistic Society of America 2021.

Gasser, Emily, Byron Ahn, Donna Jo Napoli, & Z.L. Zhou. (2019). Production, perception, and communicative goals of American newscaster speech. Language in Society 48(2): 233–259. DOI: 10.1017/s0047404518001392.

Zhou, Z.L. & Byron Ahn. (2019). Is this in the phonology? Examining the intonational phonetics-phonology interface with American English polar questions. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds), *Proceedings of the International Congress of Phonetic Sciences* 19: 2450–2454.

Faytak, Matthew, Jacob Aziz, Phillip Barnett, Jinyoung Jo, Jennifer Kuo, G. Teixeira, Joy Wu, Z.L. Zhou & Patricia Keating. (2019). Flap articulation and lowered fourth formant. Presented at Acoustical Society of America 178.

CHAPTER 1

Introduction

How do speakers organize their grammars? Though this is a defining question of modern linguistics, there are echoes of similar questions asked across disciplines. As with all scientific enterprises, linguistics can be conceptualized as fundamentally about the discovery and naming of objects and phenomena and of categories thereof. As such, classification is paramount. Classification allows scientists to place one discovery in the context of another. It gives permission to ask the questions: what else is like this (thing that I've just found)? what should I expect this to do and not do? what are its *qualities*?

Identifying a new object as like to another is thus a major part of what gives structure to scientific inquiry. The projection of past knowledge onto new is at the core of intellectual continuity and progress. Under this view, it is a *problem* when certain categories are ill-defined, as we become unable to meaningfully generate expectations and hypotheses. This dissertation attempts to address one such problem, the issue of phoneme marginality.

Traditionally, it has been said that "phonology has the categorical phenomenon of contrast at its core" (Scobbie, 2005). "Contrast" here means *phoneme*: a sound that is not other sounds, that exists apart from them and in turn defines them by being different. Scobbie's claim is historically true, for the oldest of complete phonological analyses, Pāṇini's Aṣṭādhyāyī, implicitly posits the existence of phonemes, and the American Structuralists swam in his long wake. In contrast to the phoneme is the *allophone*, a sound which, though phonetically and realizationally distinct from a second sound, is *not* treated by the speaker as different from the second but instead as a variant thereof.

But the very notion of contrast as a categorical phenomenon is and has been contentious, having been raised and revisited, criticized and critiqued, since at least 1939, when Bloomfield, one of the foremost Structuralists, observed that Menominee [uː] and [oː] only contrast in loanwords. Bloomfield said that this contrast was semi-phonemic, which I would imagine was a difficult thing to write. How can two things both be and not be the same thing? Though Bloomfield's quandary was somewhat resolved by shoving the problem elsewhere (by introducing questions about loanwords and the lexicon), there are many ways for two sounds to sort-of-but-only-sometimes contrast. Loanwords may introduce new contrasts which feel non-native to speakers, sound change can displace contrasts from one segment to another, and the same can happen due to *lack* of sound change.

These contrasts, generally not characterized as fully phonemic or fully allophonic, are called *marginal contrasts* and involve one or more *marginal phonemes* and at least one "strong", or classical, phoneme.[1] The theoretical soundness of the phoneme, *predicated* on the existence of categorical contrasts, is thus challenged by the existence of marginal ones.

Of course, there have been other, arguably more terminal, challenges to the phoneme. It has been said that the traditional phoneme had its heyday in the 1930s–1960s and has been on the decline ever since critical issues were raised by Halle and Chomsky. Yet despite this, the analytical usefulness of dividing up sounds into phonemes and allophones has never permitted these concepts their retirement. Indeed, Ladd (2006) writes, "for a theoretical construct that was discredited [...], the classical phoneme is actually still doing pretty well". Hayes (1995), writing about "foundational" concepts in phonological pedagogy, gives as his first point, "All spoken languages are phonemic in character, i.e. their segmental representations can be reduced to sequences of symbols taken from a limited inventory, namely their set of phonemes."

Outside of our field's common pedagogical practices, the phoneme has been invoked in the elucidation of many a phenomenon, from neighborhood density effects, which are calculated by using the phoneme as a unit for determining Levenshtein distance; to the categorical perception found

[1] The phenomenon of not-quite-full contrast has received many names! Hall (2013) lists 18 different terms of art, but in the general case, I will use marginal contrast/phoneme for these and strong/full contrast/phoneme for the others.

in classification tasks, where the existence of a phoneme is used to justify categoricity. Phonemic status has been invoked in explaining why within-category (phonemic category) discrimination is typically poor — speakers have little need to distinguish between different variants of a single category — and why nondiscriminability is greater if the allophonic environment is included in the task (Peperkamp et al., 2003). It has explained differences in crosslinguistic just-noticeable differences (JNDs): the JND of a given phonetic dimension is smaller for speakers which use that dimension for a phonemic contrast (Kreiman et al., 2010; Jongman et al., 2017).[2] It has been used to explain the priming of a phoneme by its allophones (Luce et al., 2003).

We should find this state of affairs alarming. If the explanations for these phenomena are fully correct, and I do not claim that they are incorrect, what do classification studies, discrimination studies, priming studies predict for marginal contrasts? Do they count as phonemes or not?

It appears that no one knows. Though probably all phoneticians and phonologists have encountered the concept of a marginal contrast, empirical investigation of category-related phenomena with respect to marginal contrasts has been rare.

Perhaps this is to be expected. After all, marginal phonemes come from a variety of sources, and it remains unclear if all are alike. In her review article, Hall (2013) gives six categories of marginal contrast grouped by origin: e.g., one class of marginal contrasts results from “distinct strata of languages” — one example being Japanese’s loosened phonotactic restrictions on borrowed words — and another from “derived contrasts” — such as opaque raising in Canadian English. These two examples are clearly genetically different, but does that mean they should have theoretically and empirically different statuses?

Consider the phenomenon of categorical perception. The received justification for its existence is that categories must contrast with each other; thus, when traversing a phonetic continuum that spans two categories, a sharp swing in categorization on the way from one category to another is reasonable and intuitive. However, is categorical perception predicted for a marginal contrast?

[2] I note that JND experiments constitute a manner of discrimination task.

Competing hypotheses can be constructed: either some amount of contrast is sufficient to cause categorical perception, so the answer is yes; or some amount of contrast is sufficient to partially sway the listener towards categorical perception, so the answer is somewhat; or only full contrast can cause categorical perception, so the answer is no.[3]

Looking to theory for answers, we find few researchers have comprehensively tackled the question of what predictions can be made regarding how category effects affect marginal phonemes or marginal contrasts. The most in-depth theoretical treatment of marginal contrasts in the literature may be Kager's (2008) demonstration that the addition of OO-correspondence and lexically specified allomorphs to classical OT results in a typology that includes "neutrast", which is a type of distributionally-limited contrast — in this way, Kager offers a treatment of one type of marginal contrast. The most expansive may be Goldsmith's (1995) argument for a five-degree cline of contrast, ranging from complete complementary distribution to complete overlap (read: contrastive in all environments). However, as Goldsmith's theory is given in about three pages in the introduction to a book, his is not so elaborated as to make predictions *per se*.

To recap, we are in an unenviable position. The root issue: a marginal phoneme is presumed to be a type of phoneme even though the concept of phoneme doesn't allow for this. We therefore lack a self-consistent understanding of a basal concept of our field. A closely related problem: it is well-documented that the existence of categories has numerous ramifications for the perceptual system but the field has only examined well-defined categories and not marginal ones. And a complication: it is generally believed that not all marginal phonemes are alike, so they cannot be treated as a group, but it is unclear how they may differentiated besides the basis of etiology.[4]

[3] Though I will be exploring this precise issue in this dissertation, there are a great many questions regarding categorical perception of marginal phonemes that I will not be able to explore. Consider that languages commonly have special "foreign" sounds which only occur in borrowed vocabulary and so these borrowed sounds will contrast with native ones only in a small subset, a single stratum, of the language. In these cases, is classification categorical always? or only somewhat? What might it depend on? Note that these questions cannot even be asked of a derived contrast marginal phoneme.

[4] That is, though Hall's review article is wide-ranging in scope, she groups her cases by how each case fails to fully adhere to some criteria of contrastiveness and not in terms of what the grouping imply for the individual languages. Some cases, failing multiple criteria, appear in multiple groups. Worse yet, cases in the same group can be very different

Of course, if theoretical progress is not forthcoming, then experimental results will have to come to the rescue: if we know how strong phonemes are affected by category effects and theory does not say how marginal phonemes will be, then we should go ahead and test the behavior of marginal phonemes. The results should be, at minimum, fertile grounds for future theorycrafting and experimentation.

In this dissertation, I present the background, methods, and results of a study on the categorical ("categorical") perception of a marginal phoneme, Canadian English [ʌ͡i]. First, in chapter 2, I summarize Hall's typology of marginal contrasts. Her work, which groups marginal contrasts by which part of the definition of phoneme is unmet, serves as the backdrop for my own classification scheme, which foregrounds general concerns about phone frequency and predictability over concerns about cause. I make the case that the lexicon is implicated in most issues involving marginality and, having done so, I then present pre-existing literature which suggests a link between lexical statistics and the strength of a contrast. I operationalize frequency and predictability as token frequency and type informativity respectively and posit that these two lexical metrics will predict participant behavior in a categorization study. Throughout, I make the case that marginality is best considered at the level of the individual instead of at the level of the language.

Thus, taking seriously the idea that phonemehood is epiphenomenal of the contents of individual lexicons, in chapter 3, I present the experimental methods behind an experiment which tests if there is indeed a connection between lexical statistics and phoneme strength. I report on the setup of a tripartite eye tracking 2-alternative forced choice categorization experiment where native speakers of Canadian English categorized three separate phonetic continua: [a͡ɪ]-[ʌ͡i], [a͡ɪ]-[ɔ͡ɪ], [ʌ͡i]-[ɔ͡ɪ]. I discuss how lexical measures were calculated on a per phone, per participant basis and how they were combined to represent a contrast instead of a sound.

This experiment resulted in two related but different datasets, which I analyze and discuss in

— should the influence of loanwords on Japanese phonology be similar to that of North Saami's /ʰr̥/, found in exactly three non-onomatopoeic words? This is not a slight on Hall's scholarship, merely an acknowledgment that looking for necessary and sufficient criteria for differentiating between cases is nontrivial.

separate chapters. In chapter 4, I present the eye tracking results by first analyzing how participants diverge in looking behavior across time, steps, and continua, finding that participants required more time to start to categorize [a͡ɪ]-[ʌ͡i] stimuli than stimuli from the other two continua and that participants have some amount of unconscious tendency to classify tokens of [ʌ͡i] as [a͡ɪ] and not as [ʌ͡i]. I also use model comparison to see if the lexical statistics successfully predict bias in eye tracking. I find that while there is a statistically significant interaction between both (summed token) informativity and continuum step and time and (type) frequency (difference) and step and time, the partial effects of these interactions are uninterpretable and the benefit to overall model fit is piddling. I conclude that, though participants treat the three continua differently from each other, it is unclear what predicts how participants differ from each other for a given continuum.

Then, in chapter 5, I turn to the categorization data. Qualitatively, it is clear that participants fall into different behavioral groups with respect to each continuum, and I discuss in turn a number of patterns which recur across continua, chiefly a reluctance to categorize stimuli as [ʌ͡i] and the observance of four varieties of categorization in the [a͡ɪ]-[ʌ͡i] continuum. Quantitatively, however, I find that these behavioral patterns generally are not predicted by a statistically significant difference in any of the lexical metrics. I discuss why this may be and suggest that the lexical metrics I used may be flawed because the calculations necessary to derive them themselves assume strong phonemehood of a marginal phoneme. Despite this, the obvious presence of behavioral groups suggests that group membership ought to be predictable by something. Moreover, the variance in behavior within behavioral groups presents another angle of attack; I end by suggesting research questions for future scholars as vexed by the problem of marginality as I.

It is clear that there is a need for us to understand the nature of marginal phonemes and contrasts. To return to my first paragraphs on the nature of science, we simply do not understand the qualities of a key component of our theories. The field has already taken for granted that there is no harm in not addressing this inconsistency — and perhaps it is right — but it seems reckless to continue without at least attempting a resolution.

When marginality has been discussed in most past research, its existence has been akin to a warning message and not an error. Not infrequently, the presence of marginality in an analysis triggers notes that marginal contrasts complicate the process of phonemicization — but not much more.[5] This dissertation says slightly more, that marginality is interesting when empirically investigated. I hope that additional perspectives on marginal phonemes and further experimental examination thereof will allow us to one day better answer the question, “what makes a contrast marginal?”, thereby allowing us to better answer its obverse: “what makes a contrast contrastive?”

[5] Scobbie & Stuart-Smith (2008) say “an unequivocal phonological system cannot be determined empirically from equivocal data”, which I believe simply reiterates that marginal contrasts are trouble.

CHAPTER 2

Background

In this chapter, I provide necessary background for the contextualization of my study, touching on three separate threads.

First, I discuss Kathleen Currie Hall's 2013 typology of intermediate phonological relationships. Of the cases that are relevant to this dissertation, I discuss how marginal contrasts typically involve low frequency marginal phonemes and/or marginal phonemes with largely predictable distributions. I propose that as these unifying factors are gradient, their influence on behavior should be likewise.

I then explain my reasoning behind choosing a categorization experiment as my probe into phoneme strength: given past results from categorization studies, the only sensible interpretation of phonemehood is one that is gradient. Armed with pre-existing results suggesting that categorization interacts with phonemic strength, it is better to continue on this path of inquiry than to strike out into the wilderness of a new paradigm.

Finally, I briefly discuss why the Canadian English [ʌ͡i] marginal phoneme is an ideal candidate for investigating how lexical factors impact contrast strength.

2.1 The typology of marginal contrasts

To reiterate, the core problem that motivates this dissertation is that phonemicity is held to be a binary proposition and marginality is neither of the options. Two paths towards a resolution present themselves. The first possibility is that marginal phonemes are simply not categories; of course this is extremely unlikely — it is obvious that if this were true, no one would have ever bothered dreaming up the idea of a marginal contrast. The alternative is that phonemicity is a gradient property, emergent from other gradient properties. If this is so, from what does phonemicity arise?

To answer this question, we will need to take a detour through what has been said about the origins of marginal contrasts and phonemes. To that end, I give a brief overview of Hall's (2013) typology of marginal contrasts wherein she gives a historical perspective on "intermediate phonological relationships" — i.e., contrasts which are not "full". Though not much has been said about how marginal contrasts and marginal phonemes may differ based on which of the criteria they fail, to say nothing of what differences we expect to see within a single group of marginal contrasts or across all marginal contrasts, I will attempt to do so by operationalizing the ideas embedded in Hall's (2013) typology.

We will arrive at the conclusion that, ultimately, contrasts are generally held to be marginal when they lack a robust enough presence in the lexicon. I thus posit that lexical statistics can be leveraged to measure the strength of a marginal contrast.

2.1.1 A typology of differences

2.1.1.1 Predictability of distribution

One classical definition of contrastiveness is that two sounds contrast if they can both occur in the exact same phonological environment. This category encompasses those cases where two sounds contrast only sometimes, e.g., only in certain environments. Hall further subsets this category into two subcategories: ones for which lack of contrast is the result of positional neutralization (and are "mostly unpredictable") and those which merely have few minimal pairs (which are "mostly predictable").

Positional neutralization marginal contrasts are of interest because few would argue that these contrasts are truly marginal. Indeed, the name of the category itself presupposes contrast. These "marginal" contrasts include Spanish /r/-/ɾ/, which contrast only intervocalically; Metropolitan

French /e, o, ø/-/ɛ, ɔ, œ/, which contrast only word-finally; and, for some varieties of Italian, /e, o/-/ɛ, ɔ/, which contrast in stressed syllables.

Hall concurs that few would assert these contrasts to be marginal. The French *loi de position* is riddled with systemic exceptions taught to students the world over. And a phonological inventory of Spanish which does not include /r/ must resort to abstraction if it is to be successful. Furthermore, the usefulness of this category is somewhat dubious, as positional neutralization is hardly a rare phenomenon. Indeed, probably every language in the world has some sort of neutralization process. To my knowledge, no one has proposed that English /s/ and /z/ are not fully contrastive just because English allows *Gatz* [gæts] but not *[gætz] and *Hodes* [ho͡ʊdz] but not *[ho͡ʊds]. As such, this category of phonological relationship is not really marginal at all.

Few-minimal-pairs marginal contrasts are much closer to what the average linguist has in mind when they imagine a prototypical marginal contrast. These include Canadian English [a͡ɪ]-[ʌ͡i], American English [l]-[ɫ], Korean [l]-[n],[1] and German [ç]-[x]. The name of this group gives it all away: marginal contrasts in this category have few minimal pairs and thus lack strong lexical support.

In opposition to positional neutralization marginal contrasts, the circumstances which give rise to this class of marginal contrast are not as commonplace. In the most interesting cases, the marginal contrast is the result of morphology (i.e., shifted contrast) and thus there is the possibility of alternations that give the speaker a hint about phonemicity. Famously, there is the case of *rider*-*writer* in North American English varieties with raising (particularly Canadian English): whereas the quality of diphthong is clearly allophonic in the unsuffixed forms *write* [ɹʌ͡it] and *ride* [ɹa͡ɪd], the effect of tapping in the suffixed forms causes the voicing contrast to shift from being redundantly evinced on both vowel quality and presence of voicing to vowel quality alone — *writer* [ɹʌ͡iɾɚ] against *rider* [ɹa͡ɪɾɚ].

Additionally, in other cases, idiosyncratic pronunciation shifts have created a budding, non-

[1] Though only in onsets — this contrast is well attested in codas.

morphologically induced contrast. Korean and German have both seen the introduction of loan words which place the "allophonic" member of the contrast in an otherwise illicit position. For Canadian English, there exist cases of both exceptional application of Raising and exceptional lack thereof. Vance (1987), for example, mentions words like *cider* [sʌ͡iɾɚ] and *fire* [fʌ͡iɚ], but also *nice* [na͡ɪs]: in these words, vowel quality is not predictable from a following consonant's voicing.

We thus see that these marginal contrasts are marginal primarily because of predictability and frequency issues. Though distribution may be primarily predictable, some number of minimal pairs exist. A minimal pair may only seem contrastive at a surface level and not when engaging more abstractly with the phonological system, but its mere existence injects ambiguity into the analysis.

2.1.1.2 Distinct strata of languages in lexicons

This category exists because languages can and do categorize their lexicons into native and foreign strata, with potentially finer distinctions drawn to separate out different foreign strata. Though there is no classical definition of contrastiveness that bears on this matter, the fact that these different strata seemingly have different phonotactics has historically been cause to wonder how much one stratum's phonemicity of contrast holds true for other strata. The gambit of "said contrast obtains only in X stratum" is a well-known one.

The most worked-out, complex, and intricate case of this is perhaps Japanese, which both stratally organizes its lexicon and has also imported a large number of new phone sequences (Itô & Mester, 1993, 2017). These include ones that disobey native phonotactics, e.g. [ti] and [ʃe], as well as ones that extend legal contrasts to new natural classes, like the extension of the singleton-geminate contrast to voiced obstruents as well as voiceless ones.[2]

Compared to the marginal contrasts discussed in §2.1.1.1, these marginal contrasts may have strong lexical support, but if so, only in a subset of the lexicon and typically in those strata which are

[2] I note that I am unfamiliar with theories of lexical strata which distinguish stratum diacritics from other lexical diacritics. This fact could be interpreted to mean that exceptionally contrastive words *in general* might be understood as merely being in a separate lexical stratum.

viewed as “non-core”. Stratal issues are quite interesting and stratum-bound phonemes naturally lead to the question of how a listener will behave when presented with the same marginal phoneme in words that imply membership in different lexical strata — sadly, a question beyond the scope of this dissertation.

2.1.1.3 Variability

Classical phonemic theory was typically unable to account for variability across the population and within the individual. In cases where speakers do vary among themselves, theoreticians often retreated to the position that the phoneme(s) in question were marginal. In my view, that this is a concern is because the field is sometimes overeager to describe languages as homogeneous systems and not a collection of similar but truly distinct idiolects.

It is debatable how many cases of this there truly are. For example, Hall includes Scottish [x], which has a “propensity” to merge with /k/ among many speakers. Another case would be the American English /w/-/ʍ/ contrast, as most speakers who have a phonemic /ʍ/ have [w] as an allophone of /ʍ/ in fast speech; further, many speakers simply do not have a /ʍ/. However, speakers with /ʍ/ are generally secure in their knowledge of which words [ʍ] and which words don’t, no matter how occasionally they realize voicelessness on those words. On the level of individuals, then, the distribution is not complex: speakers either do or do not know which words contain a particular sound. I thus will not address this category further, though that scholars have been concerned with these marginal contrasts reinforces that unpredictability of realization suggests marginality of contrast.

2.1.1.4 Frequency

Contemporary linguistic research has found time and again that frequency matters in production (Gahl, 2008; Gahl et al., 2012; Lohman, 2018) and acquisition (Maye & Gerken, 2000; Anderson et al., 2003; Thiessen & Pavlik Jr, 2016, see also Werker et al., 2012 for a brief overview of distributional learning, a token-based statistical approach to acquisition).[3] While classical phonemic analysis holds that a single minimal pair constitutes phonemic contrast, the preceding section shows that, in practice, linguists are not fully on board with this one and done approach. Yet, some languages have phones that are extremely limited in the number of words they occur in but still appear to exist as a distinct object within the mind of the speaker. Some marginal contrasts here include Arabic [lˤ], which is used "almost exclusively" in *Allah* [ʔalˤˈlˤaːh] and related words; Cairene Arabic [q], found mostly in religious words; English [x], a sound which I venture to be well-known if not well-produced by most college-educated Americans; and perhaps Korean high tone on [il], a syllable which, being vowel-initial, is predicted to have low tone by the standard analysis (Jun & Cha, 2015).

Though frequency issues do arise from different root causes — religious words, foreign borrowings, spontaneous lexical reentry; indeed, a more traditional account would probably separate out Arabic [lˤ] from English [x] — I list these cases together to highlight that a rare phoneme is rare irrespective of why. I will return to this point in §2.1.2.

2.1.1.5 Other categories

Hall addresses two more categories, "Subsets of natural classes" and "issues of phonetic similarity", which stem from totally different definitions of "marginal". One is concerned explicitly with featural distinctiveness, and the other, perceptual distinctiveness. These issues lie well beyond the scope of this dissertation so I will not treat them here.

2.1.2 Unifying factors in Hall's typology

Most previous work on marginal phonemes seeks to draw distinctions rather than create commonalities. However, this does not help with the project of understanding marginal phonemes and

[3] Matthew Faytak once shared an amusing anecdote with me concerning this. At a conference workshop on the impact of frequency on phonetics, the organizers closed the session by asking, "So what does frequency affect? Turns out, everything."

marginal contrasts as a unified phenomenon. Though this statement may seem obvious, it is illustrative to consider Cairene Arabic at this time to belabor the point.

Per Watson (2002), Cairene Arabic is replete with marginal phonemes, having eight. Two, /lˤ/ and /q/, are near exclusively found in religious words. /rˤ/ is found “predominantly in European loans” but also occurs in some native words. /ʒ/ occurs only in loans, and both /p/ and /v/ are found in “a few loan words among educated speakers”. Finally, Watson mentions /bˤ/ and /mˤ/ as rare but offers no further commentary.

How many discrete sources should we consider the above cases to come from? It is clear that the answer depends on how fine the lines are to be. One scholar may decide to group /lˤ/, /q/, and /rˤ/ as “native but marginal”, where another may only consider /lˤ/ and /q/ as such. A third may conclude that /lˤ/ and /q/ were actually (re)borrowed from neighbors and so group together /lˤ/, /q/, and /rˤ/ for a wholly different reason than the first. Does this exercise gain us anything? That is, are there substantial predictions to be made about how these marginal phonemes will behave differently on the basis of their origin?

As I stated earlier, there exist few in-depth theoretical treatments of marginal contrasts in the literature and I am aware of none which make predictions about differential effects of origin on behavior. At this moment in our field, focusing attention on the differences between cases is not nearly as productive as seeking commonalities.

In Hall’s (2013) typology, we see that the notions of predictability and frequency play important roles in if a given phoneme is felt to be marginal. Crucially, predictability is a property of the lexicon, as is frequency. In the following sections, I begin by presenting work which makes the point that lexical factors play a role in the sharpening of phonological categories through the end of adolescence. I suggest that this means we can estimate the strength of an individual’s phonological categories by examining a snapshot of their lexicon. That is, we can quantify the strength of a marginal phoneme by calculating its predictability and frequency within a speaker’s lexicon, an idea central to the experiment detailed in chapter 3. I finish by discussing how to operationalize frequency and predictability.

2.1.2.1 The implication of the lexicon

There is a body of evidence that speech perception ability is not settled in infancy but rather continues to develop through childhood and adolescence (Slawinski & Fitzgerald, 1998; Hazan & Barrett, 2000; McMurray et al., 2018, inter alia). Slawinski & Fitzgerald (1998) found that Canadian English-speaking children become more adult-like in their categorization of a /ɹ-w/ continuum through at least the age of 5. Examining British English-speaking children, Hazan & Barrett (2000) extend this finding to four contrasts (/g/-/k/, /d/-/g/, /s/-/z/ and /s/-/ʃ/) and children between the ages of 6 and 12. Finally, McMurray et al. (2018) examine the /s/-/ʃ/ and /p/-/b/ contrasts in American English-speaking children and adolescents between the ages of 7 and 18 and find the same. Though Hazan & Barrett put it more cautiously in 2000, by 2018 there was already much support for the claim that "adult-like phonemic categorization is achieved only beyond early childhood, especially for consonants".

Given this, we ought to conclude that language experience has some real effect on phonemic categorization. Since it is well-established that infants engage in distributional learning to acquire their first sound categories, this is not terribly surprising. However, the fact that development continues steadily and stolidly well into late adolescence means that special mechanisms (e.g., critical periods) are not necessary to explain development. Rather, a more parsimonious account is that natural language development through adulthood leads to stronger phonemic categorization abilities. As McMurray et al. point out, there are not that many avenues of natural language development past childhood. Of the ones which they consider, the only wholly linguistic one is lexical growth. To wit, "Children learn thousands of words during school age years. This dramatic growth of the lexicon could alter speech perception. A larger lexica (sic) could force changes in lexical competition to help make lexical access more efficient. At a fine grained level, the need to distinguish so many words *could put pressure on the system to more precisely specify phonological categories*" (McMurray et al., 2018; emphasis mine).[4]

[4] The precise mechanism of how a larger lexicon would cause this to happen is interesting but beyond the scope of this dissertation. I do note that McMurray et al. (2018) argue that "it is unlikely that further input after age 7 is

In other words, acquisition of any sound category requires experiential evidence of that sound. The literature shows we continue to "acquire" (refine) these categories through at least early adulthood. Since refinement is continuous and incremental, a parsimonious explanation is that whatever leads to refinement is both continuous and incremental; this implicates the contents of a speaker's lexicon, as it is both informed by experience and is continuously and incrementally changed over time.[5] Moreover, it is both self-evident and an empirical finding that individuals are exposed to different distributions of words.[6] Given all of this, it appears that, plainly, lexical factors should be predictors of the strength of a contrast *at an individual level, through early adulthood*. The obvious place to start is with frequency and predictability, two concepts at the heart of the problem.

2.1.2.2 Frequency measures

Every phone has some number of words which it occurs in (type frequency) and those words themselves have token frequencies. Thus, so does every phoneme and marginal phoneme.

If speakers strengthen the boundaries of their phonological categories when they encounter more unique words (types) or the same words more frequently (tokens), how many tokens (or types) must speakers be exposed to for them to build a maximally robust category? It has been conventional wisdom in the literature that token frequency is a superior predictor of frequency-related effects than type frequency for phonetic phenomena,[7] thus for the sake of managing the

necessary for learning" and propose that refinement comes through the learned ability to manage ambiguity in the signal. Despite this, their results are not *per se* in favor of this interpretation.

[5] An alternative is that it is instead neurological development that leads to refinement, though given that individual lexica are never truly fixed, I believe it most likely that both have some influence in the development of phonological systems.

[6] Both the original word gap study (Hart & Risley, 1995) and studies thereafter (Sperry et al., 2019) have found this, though in different directions. A recent meta-analysis (Dailey & Bergelson, 2022) supports Hart & Risley's (1995) result, though with a serious caveat.

[7] I acknowledge that this may be a somewhat controversial statement, but see Maye & Gerken (2000), Anderson et al. (2003), Gahl (2008), Gahl et al. (2012), Thiessen & Pavlik Jr (2016), and Lohman (2018) for examples of token-based accounts of production and perception/acquisition. It seems to me that phonological research is more inclined towards type frequency — see Bybee (1995), Albright & Hayes (2003) and Hay et al. (2004), with the latter two explicitly stating that type frequency outperforms token frequency as a predictor for their results. See also Mayer

	high token	low token
high type	Canadian [ʌ͡i], American [ɫ]	Japanese [ɸ]
low type	Arabic [lˤ], Korean [íl], Japanese [dd]	English [x], Japanese [bb]

Table 2.1: Some examples of marginal phonemes, grouped by type and token frequency.

scope of this dissertation, I will not discuss type frequency any further.[8]

Frequency measures can be calculated at the level of the language or at the level of the individual. For a particular marginal contrast, an accurate corpus is all that is necessary for calculating type and token frequencies. However, language-level frequency measures are only an approximation of the minds of individual speakers and I have already raised the point in §2.1.1.3 that an individual American English speaker has a specific lexicon while "the American English lexicon" is but a generalization across individual lexicons.

Thus I propose that we should be more interested in calculating frequency measures for individuals than for languages. To do this, in addition to a corpus, we must probe the personal lexicons of individuals by some supplementary method. For example, Canadian [ʌ͡i] is exceptionally found in a number of words which do not support its status as an allophone; speakers can be asked about these words to find the personal distribution of [ʌ͡i] for each. The corpus can then be modified such that those words have [ʌ͡i] instead of [a͡ɪ] and then frequencies can be straightforwardly calculated.

It is my belief that the prediction that contrast strength tracks frequency should hold true for individuals first and foremost, and only secondarily for a population. The strength of a marginal contrast in a language, described as a single value, can only be viewed as the aggregate measure of

(2020), which presents an algorithm for learning phonological classes with only reference to types

[8] Though I do note that marginal phonemes with low type frequency should have generally low predictability, so it may be interesting to see what happens to marginal phonemes with high low type frequency and high predictability — as it is, type frequency is likely inversely correlated with predictability to an extent that bodes poorly for model comparison purposes. That is, a model with both type frequency and a general predictability measure may find the two sharing explanatory power in a way that is problematic to detangle.

the strength of that contrast across a speaker population.

2.1.2.3 Predictability measures

Previous work on marginal contrasts has considered the influence of predictability on contrast strength. A few different ways of quantifying predictability have been influential over the years, but most measures correlate with each other. This has led to a slow change from linguistics-specific operationalizations to a generalized, information theoretic perspective. For example, one of the simplest reckonings of a contrast's predictability is functional load. Traditionally, this was calculated by counting up the number of minimal pairs distinguished by a given contrast, but a more recent method instead calculates the change in entropy in a system that would result from a given merger (Hall et al., 2021).

Two closely related information-theoretic ideas, *entropy* and *informativity/suprisal*, have been recently used in calculating linguistic predictability. Entropy is the average amount of information per event. Considering two phones, the entropy of both phones is low (exactly 0) if they are in completely complementary distribution because the event of identifying the phone provides no information that contextual information doesn't already provide. Hall (2009) presents a calculation for the conditional entropy of a contrast, a procedure which determines the entropy of a pair of sounds in a set of environments and then weights that value by the probability of each environment. That is, she calculates weighted conditional entropy.

Surprisal is the amount of information gained by an event, and informativity is what Cohen Priva (2015) calls weighted conditional surprisal. Formally, Cohen Priva defines informativity as "the weighted average of the negative log predictability of all the occurrences of a segment". That is, it is the average amount of information that a segment provides within a corpus weighted by the probability of the context. Informativity is thus low for classical allophones, as they carry no information that is not encoded elsewhere in the word and high for classical phonemes.

These two ideas are evidently closely related and, as all predictability measures can be straightforwardly determined for individuals using the same information we will collect for frequency calculations, there is no extrinsic reason to select one over the other. However, as entropy is ultimately average surprisal, and informativity is weighted conditional surprisal, I believe it makes the most sense to consider informativity as the measure of predictability which hews closest to the vagaries of the data. As such, informativity will be the measure used as a predictor.

2.1.2.4 Stratum issues

The question of how lexical strata interact with marginal contrasts is orthogonal to the other preceding issues. Though an interesting topic, I will put it aside out of consideration for scope.[9]

2.1.3 Interim summary

Grouping marginal contrasts by the origins of their marginality highlights their differences, but we are looking for similarities. We found that two lexical properties, frequency and predictability, were implicated in most marginal contrasts. How do these these properties relate to the strength of a marginal contrast? However they do, it is clear that, these measures being gradient, their influence would likewise be gradient, resulting in gradient strength across contrasts of all types.

2.2 Operationalizing strength

Even the classical definition of phoneme fundamentally relies on an appeal to the contents of the lexicon. That is, the property of phonemehood is epiphenomenal of the lexicon. Taking up this broader statement, we must imagine that the lexicon has some effect on all experimental paradigms which have been claimed to be explained by speaker phonemic inventories. Here, I discuss catego-

[9] Lexical strata *do* complicate the picture: plausibly, informativity may be high in the borrowed stratum and zero in the native. It's not clear to me what exactly this would mean, though I suspect that the idea of an "average" level of contrast across strata will not be coherent enough to make meaningful predictions. As suggested in an earlier footnote, a study considering the interaction of stratum priming with lexical factors may be valuable in clearing the air.

rization experiments as one such paradigm, highlighting its suitability as a probe into the strength of marginal contrasts.

2.2.1 The parameters of a categorization experiment

In a categorization experiment, a phonetic dimension is identified that characterizes a contrast. Stimuli are created which vary continuously across this dimension and then presented to participants who are tasked with classifying a stimulus as belonging to one of the categories embodied by the endpoints of the continuum. These experiments typically find that response curves for phonemic contrasts display categorical behavior, called categorical perception. Liberman et al. (1957) describes this phenomenon as the presence of “sharp inflections in the discrimination functions”, and informally, we can consider it as a categorization function where participant response swings sharply from one category to another at a particular point in the continuum. This is in contrast to continuous perception, where judgment changes gradually as the stimuli become more extreme. What does the shape of the response curve reflect in the mind of the participant?

A response curve can be described with a logistic equation

$$y = a + \frac{d-a}{1+e^{-g(x-x_0)}}$$

where a is the value of the lower asymptote (minimum value of the curve), d is the value of the upper asymptote (maximum value), x_0 is the inflection point, and g is the Hills slope or slope factor — not the slope at the inflection point. However, for convenience, I will use k (which does not appear in the equation) to refer to the slope of a logistic at its inflection point for the rest of this dissertation. The value of x_0, the inflection point, may vary even for very similar contrasts so its exact value is not very important. However, the exact values of a, d, and k all conceivably reflect the strength of a contrast.

Imagine a language with a robust /p-pp/ contrast but a marginal /b-bb/ contrast. The /p-pp/ contrast, varying primarily on duration, is expected to exhibit categorical behavior: in a task where

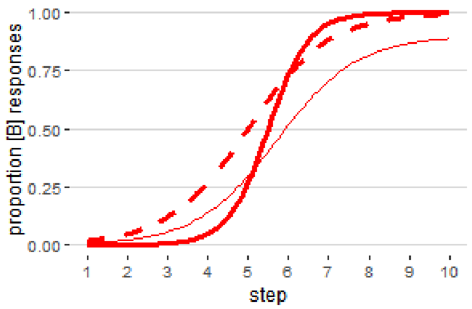


Figure 2.1: Schematic response curves representing contrasts with different degrees of categoricity. The thick, solid line is a categorical response, corresponding to a strong contrast. The thick, dashed line and the thin, solid line are possible response curves corresponding to a marginal contrast.

the participant is asked to determine the identity of a stimulus ambiguous in duration between /p/ and /pp/, they will, at some point in the range between typical durations for /p/ and /pp/, rapidly switch from identifying shorter stimuli as singleton to identifying longer stimuli as geminate. The marginal /b-bb/ contrast might then be different in at least one of the parameters that define a response curve.

One possibility is that categoricity of response is primarily quantifiable by the value of k, the steepness of the response curve. If the value of k is high, the curve is steeper and represents a more categorical response. In other words, high k results in the canonical categorical perception response curve shape and is represented by the thick, solid line in Fig. 2.1. The thick, dashed line, on the other hand, has lower k and represents a more gradient, less categorical response pattern. This is an obvious possibility for what the results of an identification task investigating a marginal contrast might look like. Given the above discussion, we can predict the steepness of the response curve to be inversely correlated with predictability. In other words, the less predictable a sound is, the steeper the response curve, the greater the value of k, the more categorical of a response curve.

An orthogonal possibility is that participants are never fully willing to identify even extreme stimuli as some category. This would be represented by a lower value for d, the maximum value of the curve, and is represented by the thin, solid line in Fig. 2.1.[10] This is a less obvious possibility for what the response pattern of a marginal contrast might look like and indeed may represent a separate type of marginality, perhaps stemming from very low frequency. Alternatively, response curves with low d may indicate some underlying unnaturalness of stimuli. In any case, d represents a sort of categorizational commitment, with low d corresponding to a reluctance to classify extreme stimuli as instances of its putative category.

2.2.1.1 Previous work

The question is then if these predictions hold water. Thankfully, the connection between the strength of a contrast and the parameters of its response curve has already been positively demonstrated by Gelbart (2005).

In his dissertation, Gelbart conducted a categorization task on native Japanese speakers, examining their willingness to identify different steps on a consonant duration continuum as either singleton /b, d, p/ or geminate /bb, dd, pp/. Crucially, geminate voiced stops are phonotactically licit only in foreign, borrowed words, so Gelbart's expectation was that speakers will give a more categorical response for the /p-pp/ continuum and a less categorical response to the /b-bb/ and /d-dd/ continua.

This is, in essence, what was found — Fig. 2.1 is inspired by one of his summary graphs, with /p-pp/ being the thick, solid line, /d-dd/ the thick, dashed line, and /b-bb/ the thin, solid one. This finding, that Japanese speakers have a more categorical perception of a /p-pp/ continuum than a /b-bb/ or /d-dd/ one, is congruent with the idea that marginal contrasts have different response curves than strong contrasts, as voiced geminates only appear in loans. Participants were also unexpectedly unwilling to characterize any token, no matter the length, as /bb/. The reason for the low d of the /b-bb/ case is not fully explained, though a similar reticence to identify stimuli as [bb] was found

[10] Alternatively, it could be a higher value for a — for convenience, I will refer only to d in this section, with the understanding that a d that is less than 1 and an a that is greater than 0 both represent unwillingness to posit some category.

by Kawahara (2005). Earlier, I intimated that this squashed response curve may be the result of insufficient lexical support; that suggestion comes from Gelbart's (2005) experiment plus the fact that Kawahara reports that /bb/ is the rarest of the geminate voiced stops in Japanese, with only about 500 tokens found in the Nihongo-no Goitokusei (Lexical properties of Japanese; Amano & Kondo, 2000) corpus versus about 23000 /dd/ tokens.[11]

2.2.1.2 Using bias to compare troublesome contrasts

Gelbart is able to directly compare the response curves for the three examined contrasts because all continua vary along the same dimension, that of duration. However, a notable pitfall of the categorization experiment is that different cases often cannot be directly compared to each other as even minor experimental decisions, such as how many intermediate steps to include, will affect the parameters of the response curve: if a particular experiment uses only stimuli that are close to a categorical boundary, the slope of the response curve will appear less steep. Thus, though comparison of response curves can be informative, direct comparison of response curve parameters requires that all experiments be normalized to something which all continua have in common, such as JND. Yet determining JND for every single continuum we might want to consider would not be time- or effort-efficient.

Additionally, it could also be informative to directly compare the response curves for two continua if the endpoints between them are phonetically similar — or better yet, if one of the endpoints is fixed in the two continua and the contrasting endpoints are phonetically similar. Such a scenario would likely involve vowels in a language with many of them, as phonetic similarity is generally harder to come by in consonantal contrasts.

To circumvent this comparison problem, this dissertation adapts the elaborated categorization task employed in McMurray et al. (2018). McMurray et al. are also interested in understanding

[11] Sadly, Gelbart did not examine the /g-gg/ contrast. In Japanese, /gg/ (about 1200 tokens) is also much less frequent than /dd/, so it would have been highly informative to see which of the /b-bb/ or /d-dd/ response curves a /g-gg/ curve would have resembled. Based on the results in Kawahara (2005), it should have looked intermediate between the /b-bb/ and /d-dd/ response curves, either in slope, maximum, or both.

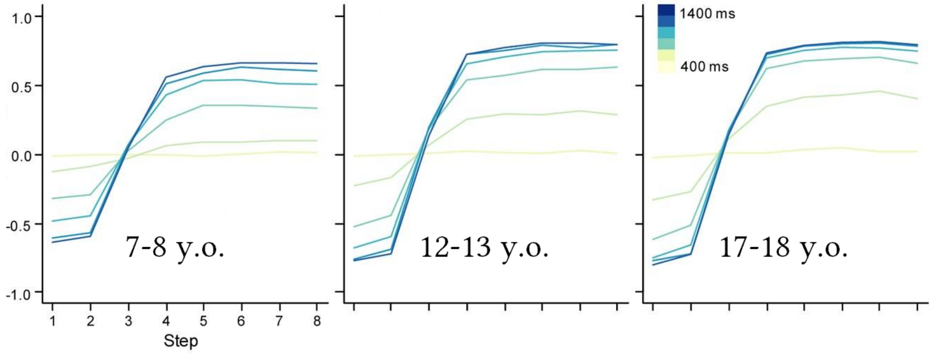


Figure 2.2: Bias at different time steps for two different phonetic continua from McMurray et al. (2018). The three panels represent different age groups and show that older participants are faster to categorize (distance between bias lines is greater for 17–18 y.o.'s than 7–8 y.o.'s) and more willing to categorize as well (more extreme values of bias are reached by 17–18 y.o.'s).

how the k and d[12] in §2.2.1 vary across time and across different continua. They, too, are interested in measuring differences in the strength of two phonological contrasts, /s-ʃ/ and /b-p/ and to do this without requiring normalization, they look directly at bias and not at response curves. In their experiment, participants had their eyes tracked during a visual categorization task. This was then used to calculate the bias (average looks to the one category minus average looks to other) at 20 ms intervals across trials, giving the graphs reproduced in Fig. 2.2.

These graphs essentially provide a timeline of how much certainty participants have that a given token is of some category at every time for each step and the authors note that the result is "something analogous to a standard identification curve". For example, in the left panel in Fig. 2.2, the lightest yellow line at t=400 ms shows a bias of essentially 0 for all steps and the next lightest line representing t=600 ms shows nonzero bias at steps not equal to 3. This can be compared directly with the same two lines (t=400 ms, t=600 ms) in the right panel of Fig. 2.2, where the bias at steps not equal to 3 is greater in absolute magnitude. This shows that, given the same amount of time, older participants reached higher levels of bias (were able to categorize faster) than younger

[12] To be precise, they are not interested in d but rather $d - a$, but this is immaterial to my discussion.

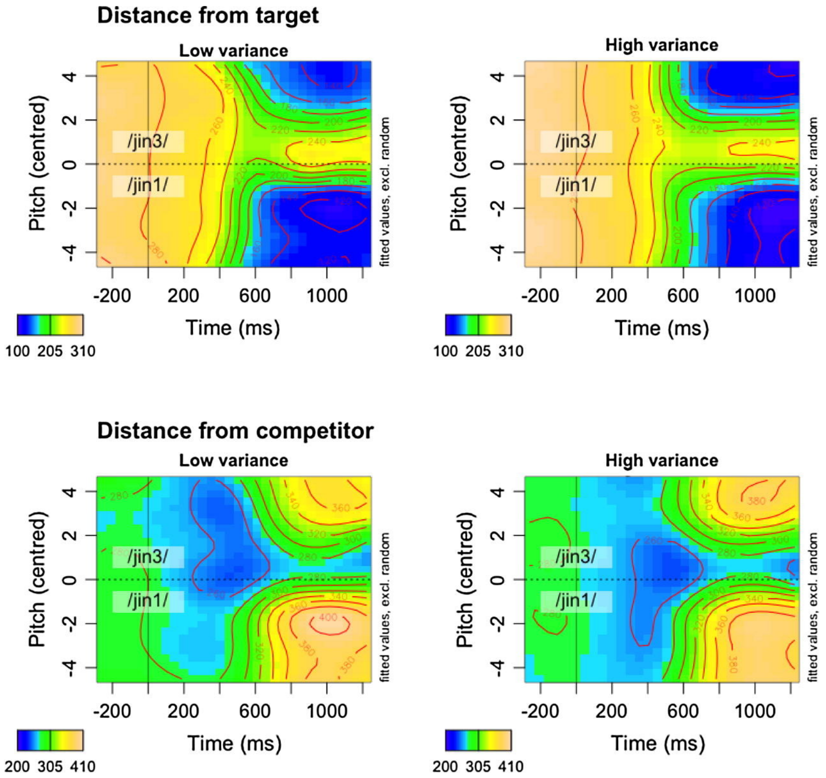


Figure 2.3: An alternate graphical representation of bias at different times across continuum steps from Nixon et al. (2016).

participants.

The data can also be used to generate evocative contour plots of the type in Fig. 2.3. Nixon et al. (2016) use a similar experimental paradigm as McMurray et al. (2018), and though their object of study is different, they also make use of bias as their dependent variable. In this figure, the colors represent bias, the continuum is shown on the y-axis, and time on the x-axis.

As these contour plots simultaneously represent two independent variables and one dependent, they have the advantage of being ideal for compactly displaying interaction effects. For example, by horizontally scanning across the top right panel in Fig. 2.3, it can be seen that bias decreases across time for both pitch=4 and pitch=-4, but that the lowest bias (here, purple) is reached by t=1000 ms at pitch=-4 while it takes about 100 ms longer at pitch=4. This result is analogous to the result described by Fig. 2.2, but the presentation allows for participant certainty to be simultaneously compared across time, step, and even continua: the extreme values of bias reached within a continuum represent degree of commitment, and the closeness and number of contour lines at a given y value represent how quickly that commitment is reached.

2.2.2 Interim summary

I have proposed that we revisit and take seriously the idea that phonemehood is epiphenomenal of the lexicon. Presently, I am of the view that predictability and frequency of a sound must have something to do with its acquisition — the historical view of minimal pairs constituting phonemehood directly implicates predictability, and work such as Anderson et al. (2003) and Thiessen & Pavlik Jr (2016) implicate frequency. This perspective implies that phonemic contrasts are not all alike in strength, so perhaps the strength of a contrast is proportional to that which affects the acquisition of said contrast. Put concretely, frequency and predictability should be proportional to the strength of a contrast if contrast can be gradient: gradient predictors, gradient results.

Examining studies in the literature which invoke the notion of categorical contrast, I argued that it has already been shown that weaker lexical support results in weaker effect strength when categoricity is invoked in explaining the strength of a phenomenon. I thus identified categorization experiments as a paradigm useful in further empirically testing this link between behavior and the lexicon.

In a categorization task which finds categorical perception in the phonetic space between [p] and [b], the (degree of) binarity of response (*viz.* k) is traditionally attributed to the fact that /p/ and /b/ are both phonemes that show a robust, phonemic contrast. The first interpretation follows: *mutatis mutandis*, an experiment looking at two phonemes with a weaker contrast should find less binarity of response; all else equal, k should be less, perhaps approaching a straight line.

Chosen carefully, a phoneme and its marginal counterpart will contrast weakly — a categorization task thereof should show a smaller k than one involving categorization of a strong contrast. This all follows from the assumption that k tracks strength of contrast. Further, the marginal member in a weak contrast may be dispreferred by participants, resulting in a less extreme upper asymptote (d) or a lower asymptote (a).

A different scholar, drawing from a classical understanding of phonemes, might say that all phonemes are the same strength (i.e. that there is no such thing as "two phonemes with a weaker contrast"), so any two strong phonemes will show a high k and everything else, an exceedingly low k. I assert that the second option has been ruled out by Gelbart (2005), as his study finds different values of k for different phonemic contrasts. The first interpretation — that k tracks contrast strength — is the only sensible one. There is otherwise no way to make sense of both the extensive literature on categorical perception *and* Gelbart's (2005) dissertation unless Gelbart's results are wrong.

If contrast strength is continuous, now understanding "phonemic contrast" as a case of very high category strength, it must be that all results which invoke phonemic contrast are fundamentally about category strength and thus lexical support. Therefore, those results will change accordingly for participants with differing lexicons and differing perceptual experience. Indeed, what we have seen thus far is that experiments which do find low k and/or low d look precisely at populations with more restricted lexicons. It is likely that the Japanese speakers in Gelbart's (2005) study did not all have the same words in their English loanword stratum,[13] and McMurray et al. (2018) directly note that their participants had developing lexicons.

Having elaborated on this chain of logic, I discussed a modification to the classical categorization task that helps with principled comparison of categorization response curves — by looking at bias instead of categorization choice. I now turn to the task of selecting a test case.

[13] In his dissertation, Gelbart notes that the twelve participants whose categorization data contributed to the response curves were all native speakers of Japanese from "among and around the U-Mass, Amherst community." In a different experiment on the same population (though not on the same participants), he (or his committee) was evidently concerned enough with the possibility of "variance that might have been due to the advanced bilingualism of some of the participants in the original running" to rerun that other experiment in Japan. Frankly, it is unclear to me if this means his response curve results are suspect, but I do maintain that it is highly unlikely that any two individuals have the exact same lexical entries relevant to a given contrast. The data I collected using the survey I describe in §3.1.4 found that none of the 31 participants included in my final analyses had a lexicon identical to that of another.

2.3 Canadian English as a test case

Because we expect marginal contrast strength to be affected by lexical statistics, and because generalizations about the lexicon of a population are necessarily approximate, an ideal test case, then, would be a marginal contrast known to have many exceptions, and, indeed, one where individuals vary on which exceptions they maintain.

Enter Canadian English which, thanks to the process of Canadian Raising, has the diphthongs [a͡ɪ, ʌ͡i, ɔ͡ɪ], the first and last of which are clearly phonemic while the middle is marginally so.

Historically, [ʌ͡i] was an allophone of /a͡ɪ/ that occurred before voiceless obstruents. However, this generalization, called Canadian Raising (CR), was made opaque by a different process, tapping, which turned /t, d/ into [ɾ] in many environments. It thus transpired that [a͡ɪ, ʌ͡i] seemed to contrast before [ɾ], though only at a surface level — through morphophonological alternations, a speaker could reconstruct the underlying form of a word, undo tapping and CR, and negate this apparent contrast. *However*, as Vance (1987) documents, [ʌ͡i] started to appear in the "wrong" environment (i.e., not before a voiceless obstruent) in some words, such as *cider* and *spider*, but also *fire*, *ire*, and *irony*. In other cases, such as *cyclops*, *icon*, and *python*, CR mysteriously failed to apply, leaving unraised diphthongs before voiceless obstruents. Notably, these exceptions vary — different speakers have different exceptions. [ʌ͡i] is in fact, famous in the profligacy of its exceptionality.[14] This state of marginal contrast has persisted in Canadian English since at least 1942, when Joos mused on the possibility of a phonemic split between raised and unraised diphthongs.[15]

Canadian English presents a perhaps unique opportunity to take a look at frequency and predictability effects using a single language. First, Canadian English is accessible, being spoken by millions.[16] Further, as the number of exceptions to CR is greater than that of the equivalent

[14] Of course, Canadian Raising also applies to /a͡ʊ/, but most past research focuses on [ʌ͡i] and we have a correspondingly more detailed understanding of exceptions involving [ʌ͡i].

[15] He is ultimately equivocal: "There is no use in guessing [what] will happen [... but] perhaps I have gone too far for the present state of our science: perhaps this sort of prediction is not legitimate".

[16] Particularly accessible to myself, a native speaker of American English on the same continent.

phenomenon in the US, individuals are potentially more variegated in their list of exceptions.

The vowel inventory of Canadian English also allows us to examine the strength of a contrast involving a marginal phoneme and a strong phoneme which it is *not* related to: the presence of [ɔ͡ɪ] permits testing of categorization in the [ʌ͡i]-[ɔ͡ɪ] continuum and comparison of those results to categorization of [a͡ɪ]-[ɔ͡ɪ]. Not only has a categorization experiment never been done on a marginal phoneme and an unrelated strong phoneme, this is exactly the situation described in §2.2.1.2 that would allow us to directly compare response curves across continua.

Thus, Canadian English is an excellent test case for this dissertation. The strength of the three contrasts between [a͡ɪ, ʌ͡i, ɔ͡ɪ] can be investigated with a categorization task. After the task, participants can be given a questionnaire on words that have been documented to have unpredictable diphthong quality and asked to identify their exceptions. These results can then be used to generate frequency and predictability measures on a per phone basis for [a͡ɪ] and [ʌ͡i][17] and then combined to create per contrast lexical metrics — a procedure I will discuss in §3.2.

Both bias and individual response curves can be compared to see whether performance is dependent on the contrast-level lexical metrics, and progress can be made on answering the question of what makes a marginal contrast more or less marginal. And indeed, if such a question is sensible in the first place.

2.4 Chapter summary

This background chapter has touched on the theoretical underpinnings of marginality, concluding that frequency and predictability are prominent considerations to if a sound is deemed marginal. I have proposed that we can operationalize frequency and predictability as token frequency and informativity and that they should be calculated at the level of the individual, not at the level of a language.

[17] For simplicity, I assume that [ɔ͡ɪ] will be consistent across speaker lexica, though of course this is not entirely accurate.

I then discussed how contrast strength has been operationalized by previous work. Looking at Gelbart (2005) and McMurray et al. (2018), though the former conducted a classical categorization experiment and the latter, a more involved paradigm with eye tracking, both projects were interested in the slope k and the asymptote d of their respective curves. While Gelbart's (2005) experiment is more straightforward, McMurray et al.'s (2018) approach has a certain advantage in sidestepping some inherent problems with comparing different phonetic continua. Still, an ideal case would allow for both methods.

That case happens to be found in Canadian English with its [a͡ɪ, ɔ͡ɪ] phonemes and its [ʌ͡i] marginal phoneme. Not only are Canadian English speakers easily accessible, the number of exceptions to the general rule is large, so the lexical support for [ʌ͡i] can vary greatly within the popu-laiton of native speakers. Furthermore, both eye tracking and categorization curve comparison can be done by having participants simply categorize more phonetic continua than just [a͡ɪ]-[ʌ͡i].

Equipped with this background, I turn now to the design and methods of the experiment at the heart of this dissertation, a categorization task on [a͡ɪ]-[ɔ͡ɪ], [a͡ɪ]-[ʌ͡i], and [ʌ͡i]-[ɔ͡ɪ].

CHAPTER 3

Experiment design and methods

In the last chapter, I presented the chain of logic that if contrastiveness is the basis of a certain phenomenon, if lexical support is the basis of contrastiveness, and if contrastiveness can be gradient, then a population with varying lexical support for a contrast should show an effect of individual lexicons on how that certain phenomenon plays out within those individuals. To test this hypothesis, I conducted an eye tracking categorization study involving one of the best-studied marginal phonemes of contemporary linguistics, Canadian English [ʌ͡i]. Here, I present the experimental design and methods of this study.

3.1 Experimental design

The core of the experiment was a categorization task spanning three phonetic continua: [a͡ɪ]-[ʌ͡i], [a͡ɪ]-[ɔ͡ɪ], and [ʌ͡i]-[ɔ͡ɪ]. Following McMurray et al. (2018), the study was an eye tracking two-alternative forced choice (2AFC) task in the visual world paradigm where participants categorized stimuli as one of [a͡ɪ, ʌ͡i, ɔ͡ɪ] while looking at images they had been trained to associate with those sounds. This experiment took place in Ashley Farris-Trimble's Phonological Processing Lab (PPL), and I am most grateful to her, her lab manager Danica Reid, and her research assistants.[1] When discussing the experimental design and the minutiae of running the experiment, all references to "we" refer to myself and plus Ashley Farris-Trimble and Danica Reid.

[1] Particularly because Ashley helped me prepare the IRB and originally agreed to let me use her facilities with the expectation that we would produce a paper! Ashley, I truly am forever grateful.

3.1.1 Participants

Participants were limited to undergraduate students attending Simon Fraser University (SFU) between the ages of 19–35. All participants were self-professed native speakers of Canadian English, as assessed by a survey. Our IRB approval did not allow us to collect age or gender information, and we did not target any particular age or gender distribution. We did not collect information about where participants grew up, nor did we ask directly if participants had Canadian Raising, though our survey results indicated that all participants did have [ʌ͡i] in at least some words — see Appendix C for a summary of participant responses to our pronunciation survey.

Research assistants from the PPL recruited participants by visiting classrooms and tabling in public areas of the university. To further incentivize participation, participants could elect to be entered in a raffle to win 1 of 7 USD$50 Amazon gift cards. At the time, this was approximately CAD$70.

A total of 47 participants took part in the study, though 16 were excluded due to technical issues, inattentiveness, and inappropriate language background. In the end, we obtained usable data from 31 participants, totaling 15.8 million eye tracking observations (though we will bin these in groups of 10 for analysis; see §3.3 for details) across 14697 categorization trials.[2]

3.1.2 Stimuli

The triplet [kla͡ɪɾɚ, klʌ͡iɾɚ, klɔ͡ɪɾɚ] were chosen as nonce words to represent the endpoints for the three continua. As shown in Table 3.1, these nonce words have similarly impoverished phonological neighborhoods, though somewhat different phonotactic probabilities, and were judged by us to not be reminiscent of actual words.

Danica, a phonetically trained, female native speaker of Canadian English, recorded these

[2] Participants 23, 30, and 31 were missing one trial each, for unknown reasons. Very mysteriously, participant 46 was missing 180 trials. As the missing trials were randomly distributed across continua and step, their data was included in the final analyses.

word	“Klattese”	N. density	N. frequency	Σ unigram probability	Σ bigram prob.
kla͡ɪɾɚ	klYdX	0	0	1.2329	1.0129
klʌ͡iɾɚ	klYtX	3	5	1.2820	1.0185
klɔ͡ɪɾɚ	klOtX	4	1.25	1.2712	1.0150

Table 3.1: Neighborhood density and frequency, and unigram and bigram probabilities, calculated by the KU calculators (Vitevitch & Luce, 2004, 2016).

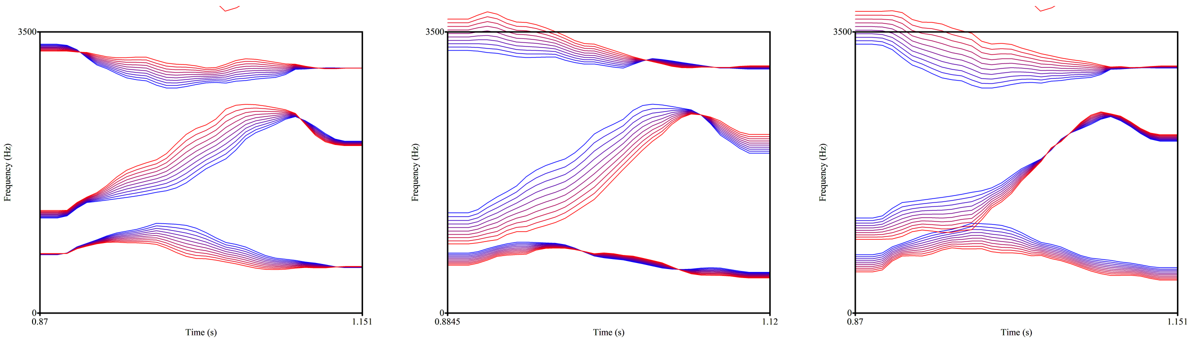


Figure 3.1: Formant trajectories of stimuli. Blue and red endpoint trajectories are original, all other trajectories calculated by Winn 2019. From left to right, [a͡ɪ] in blue – [ʌ͡i] in red, [ʌ͡i] in blue – [ɔ͡ɪ] in red, [a͡ɪ] in blue – [ɔ͡ɪ] in red

words in the frame “Click on the X” and the Make Formant Continuum Praatscript was used to synthesize three phonetic continua of ten steps each (Winn, 2019).[3] All items had an average intensity of 73 dB, nearly identical intonational contours, and formant contours are given in Fig. 3.1.

Each nonce word was randomly associated with an image of an object and participants were taught this association in a training block which occurred immediately before the eye tracking portion of the study. We used the NOUN Database (Novel Objects & Unusual Name; Horst & Hout, 2016) for the novel images. Specifically, we used images 2005, 2025, and 2054 as images with high novelty (in Horst & Hout’s (2016) norming study, only 6% of participants had ever encountered the most frequently encountered object in this trio) and low nameability (15% or fewer of participants

[3] As it is quite difficult to replicate the exact outputs of the script — Winn calls the use of this script a “dark art” — they are available in the supplementary materials at osf.io/4xveb.

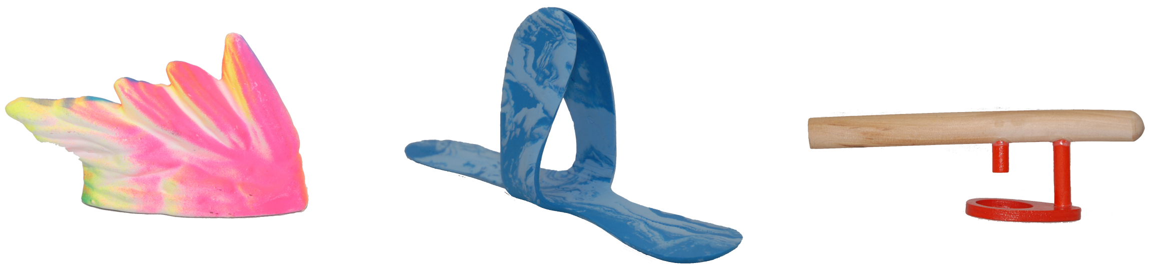


Figure 3.2: Images from the NOUN database. From left to right, images 2005, 2025, 2054.

spontaneously came up with the same name for the most nameable object in this trio).

3.1.3 Method

After obtaining informed consent, a research assistant led participants to a computer paired with an Eyelink 1000 eye tracking device (running on the v4.56 software).

3.1.3.1 Training block

The experiment began with a training block to teach participants to correctly associate images with the nonce words. We conducted a pilot study to determine that this training block would be 60 trials long; each word was the correct choice in 20 trials, 10 times with both other words. The logic for the training block was written by Danica in ExperimentBuilder (SR Research Ltd., 2020).

Before putting on headphones, participants were read the first part of the instructions by a research assistant (Appendix A, training portion).

The training block initialized by randomly pairing the three images with the nonce words [kla͡ɪɾɚ, klʌ͡iɾɚ, klɔ͡ɪɾɚ]. Participants were presented with a small black dot in the center of the screen. After clicking on the dot, participants were presented with two images on the left and right sides of the screen and a recording corresponding to one of the continuum endpoints then played over the headphones, instructing participants to "Click on the X." 25 ms after clicking, a green box appeared over the object which should have been associated with the nonce word.

3.1.3.2 Calibration procedure

After completing the 60 training trials, participants were formally introduced to the Eyelink 1000 machine and engaged in a 13-point monocular calibration. At this point, participants were read the second part of the script by a research assistant to explain the calibration procedure (Appendix A, calibration instructions).

Our standard procedure was to never do binocular calibration as we only ever planned on tracking a single eye. By default, the left eye was calibrated and then tracked, unless calibration failed twice, in which case a research assistant calibrated and tracked the right — a situation which happened with only four participants over the course of the study.

3.1.3.3 Testing (categorization) block

Once training was completed successfully, participants were read the third part of the instructions by a research assistant (Appendix A, experiment instructions) before the eye tracking portion of the study commenced. This part of the study was a single block comprising 3 continua × 10 steps per continuum × 16 repetitions per step for a total of 480 trials. The logic for the categorization block was also written by Danica in ExperimentBuilder.

In this portion of the experiment, the word-image pairings were inherited from the training block and trials were presented to participants in a totally random order.[4] The Eyelink 1000 device was set to a sampling rate of 500 Hz.

In each trial, the two objects corresponding to the endpoints of a continuum were randomly

[4] We considered a design where the eye tracking portion would be composed of three blocks, where each had only stimuli from one contrast, allowing counterbalancing of continua. We decided that random order would likely be more engaging to the participants, as the experiment was somewhat lengthy.

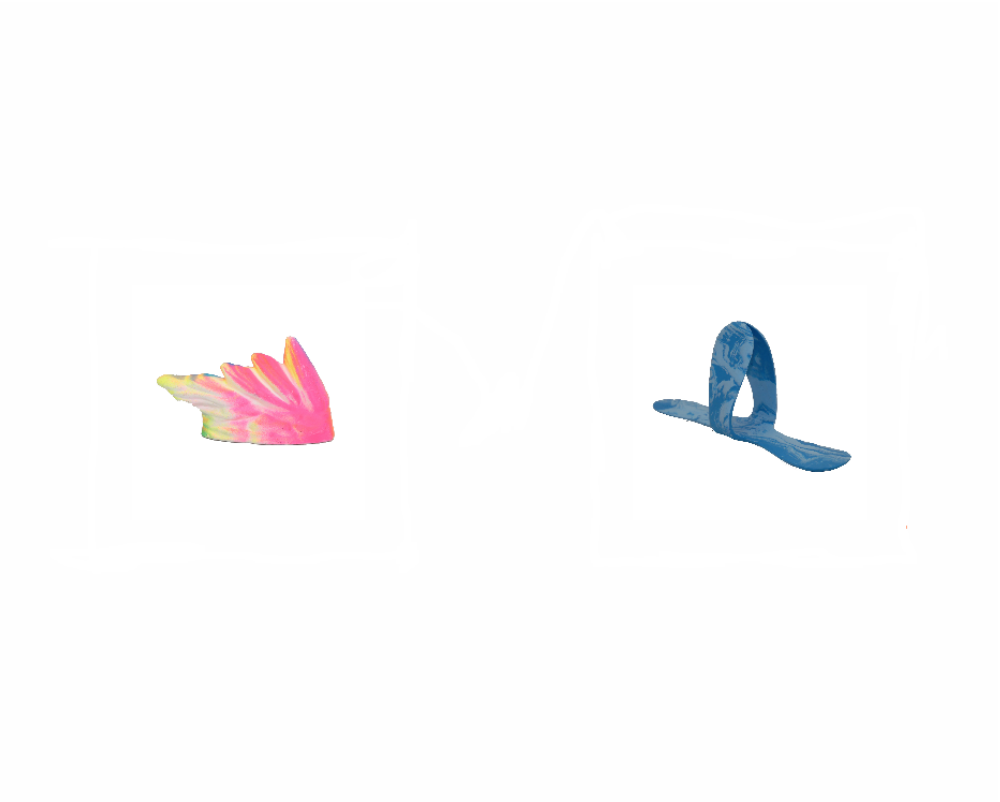


Figure 3.3: View of experiment participants saw during testing and training blocks.

placed in the left and right sides of the screen and a red dot was placed between them. Participants were instructed to look at the red dot until it became a blue dot. This happened 500 ms after the images first appeared. Participants then had to click on the blue dot, whereupon the stimulus played over the headphones after a 25 ms pause, instructing the participant to "Click on the X". At this point, the selected image was recorded by the experiment.

As 480 trials is somewhat lengthy for an experiment of this type, participants were given the opportunity to take a brief break after every 30 trials.

3.1.4 Survey

After completion of the eye tracking block, participants were asked to complete a survey about their language usage by a research assistant reading the fourth part of the instructions (Appendix A, survey portion). I used jsPsych (de Leeuw et al., 2023) to write the logic for the survey and we hosted it on a SFU JATOS server so that participants could complete it at their convenience.

In the survey, participants first self-reported on their language history before telling us how they pronounce words with potentially exceptional diphthongs, that is, words which may contain a diphthong not predicted by the allophonic rule.

First, a selection of words noted by the literature to have exceptionally (un)raised diphthongs was collected. We used Chambers (1973, 2006); Hall (2005, 2012); Idsardi (2006); Pater & Moreton (2014); Vance (1987) and Bermúdez-Otero (2003). I also solicited suggestions from linguists in the UCLA linguistics department.[5] After compiling the list of potentially exceptional items, which we called our `exception_list`, I then created an `exhaustive_exception_list` which contained all words transparently derived from words in `exception_list` as we expected these derived words to pattern along with their headwords in exceptionality. That is, if a participant has an exceptionally raised [ʌ͡i] in *dine*, then we expected that to also be true for *dines*, *dined*, and *din-*

[5] Thanks particularly to Kie Zuraw.

ing. I pared down the list to include the 51[6] headwords with the most frequently occurring derived words in the $\text{SUBTLEX}_{\text{US}}$ and used this final list, available in Appendix B, to create the survey.

We first attempted a version of the survey where we asked participants to compare a word of interest with a word with a known vowel and indicate if the two words rhymed. Though we thought that asking about rhymes might be a good way of learning which words were exceptional for a given participant, a pilot showed that something about the task was more difficult than expected, and people became confused about how exactly they said words.[7] Thus, instead, we decided a direct comparison might be better. Danica, the same native speaker who recorded the endpoint stimuli for the continua, recorded herself producing each of the words of interest with both a raised and unraised diphthong,[8] and the final survey asked the participant to listen to these reference recordings before selecting the recording which sounds most like how they would say the word. As with all other stimuli, these recordings are available in the supplementary materials at osf.io/4xveb.

The survey also had three catch questions at roughly the 30%, 60%, and 80% marks. These questions asked the participant to select only a specific option, disregarding the actual question. Missing a single catch question resulted in exclusion from analysis. A summary of responses to the survey from all participants who were not excluded from the study is available in Appendix C.

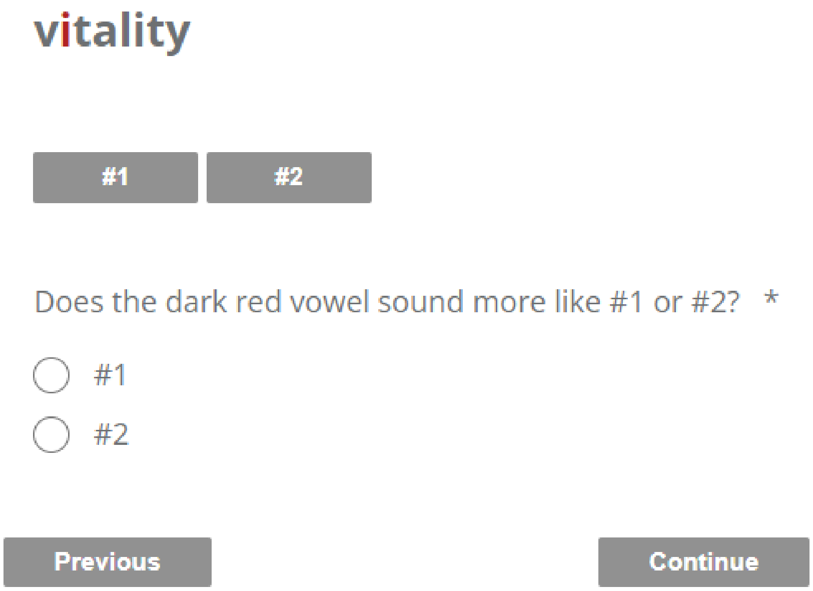
3.2 Calculation of lexical predictors

To calculate the type informativity and token frequencies of the three diphthongs [a͡ɪ], [ʌ͡i], and [ɔ͡ɪ], we needed a frequency corpus of English words and information about which vowels are in which words. We decided to use the $\text{SUBTLEX}_{\text{US}}$ corpus (Brysbaert & New, 2009) as the fre-

[6] Originally 52, but we decided to exclude *semifinal* as everyone in a pilot preferred [sɛmi-] over both [sɛma͡ɪ-] and [sɛmʌ͡i-].

[7] A story I have heard from multiple people who have an [a͡ɪ]-[ʌ͡i] distinction is that, as children, they were confused about the phrase "I scream, you scream, we all scream for ice cream" because of course it's less punny if "I scream" and "ice cream" don't rhyme. As such, it was surprising that asking about rhymes was not as intuitive to participants as we had expected.

[8] Thanks very much, Danica!



vitality

Does the dark red vowel sound more like #1 or #2? *

○ #1

○ #2

Figure 3.4: View of survey. See Appendix A, survey instructions, for how participants were asked to respond.

quency corpus; although a similar corpus (i.e., of naturalistic, spoken dialogue) based exclusively on Canadian media would have been preferable, such a corpus was not available and we judged that $\text{SUBTLEX}_{\text{US}}$ was a suitable enough approximation of word frequencies in Canadian English.[9] I used a copy of the CMU pronouncing dictionary to determine the expected pronunciation of each word in $\text{SUBTLEX}_{\text{US}}$. This was done by using Python to combine $\text{SUBTLEX}_{\text{US}}$ and CMU and writing a regular expression that captured the environment of Canadian Raising to "implement CR" within the combined dataframe which we called `CMU-SUB_raised` — that is, I created a version of CMU with accurate transcriptions that assumes perfect application of the allophonic rule. In essence, this created a corpus with phonetically accurate transcriptions of all [a͡ɪ] and [ʌ͡i], assuming a noncontrastive distribution.

Through completing the surveys, each participant had provided us with a list of their own exceptional words. I again used Python to propagate each individual participant's exceptions into `exhaustive_exception_list` before using the individual-specific list of exceptions to exceptionally raise/undo raising within `CMU-SUB_raised`, resulting in a unique `CMU-SUB_raised` corpus for each participant.

This participant-specific corpus was then loaded into Phonological Corpus Tools 1.5.1 (PCT, Hall et al., 2021). After loading a customized `.feature` file which allowed PCT to handle the existence of [ʌ͡i] — see Appendix D — I used PCT to calculate the token frequency per million words (SUBTLWF) and type informativity of [a͡ɪ], [ʌ͡i], and [ɔ͡ɪ] for each participant. These measure are all available in Appendix E.

As a worked example, the default `CMU-SUB_raised` gives a token frequency of 79711.5 (per million) for [a͡ɪ] and 17279.4 for [ʌ͡i]. Now, *fire* has a token frequency of 215 in the corpus, so a speaker who indicates that they have [fʌ͡iɚ] as their only exception to CR would have, instead, a token frequency of 79711.5−215=79496.5 for [a͡ɪ] and a token frequency of 17279.4+215=17494.4.

[9] Some Canadian corpora were considered, but most corpora were totally text based. The Strathy corpus of Canadian English (Davies, 2012), for example, touted itself as balanced between a number of different speech styles. After looking at its contents in greater detail, however, it was clear that the only spoken data comes from transcripts of town hall meetings.

	mean	sd	min	max
a͡ɪ-ɔ͡ɪ	77033	1362	75906	82244
a͡ɪ-ʌ͡i	62831	2724	60577	73254
ʌ͡i-ɔ͡ɪ	14202	1362	8990	15328

Table 3.2: Summary statistics for token frequency difference for all three contrasts. All numbers in occurrence per million words.

Having discussed in chapter 2 why I believe informativity and frequency to be worth investigating in this study, it is important to note that, so far, what has been calculated are phoneme-level metrics, not contrast-level ones. Thus, the last step here was to combine the individual lexical statistics for each phoneme into a combined measure that would hopefully be relevant to the contrast. The existing literature does not offer any guidance on how this might be done, so I had to first reason about combined measures that could not be meaningful before selecting combined measures that had the best chances of being good predictors of categorization behavior.

3.2.1 Frequency

Because of how we treated [a͡ɪ] and [ʌ͡i] in participant lexicons, any increase in token frequency for [a͡ɪ] is balanced by an identical decrease for [ʌ͡i] — that is, $R^2 = 1$ for these variables. Thus, while it is reasonable to posit that participants would react faster to a contrast where both options are frequently heard, it is not possible to test that hypothesis with the data as collected. Therefore, frequency sums cannot be informative predictors. I decided instead to calculate the difference of the frequencies, with the idea that a participant with a frequency distribution more skewed towards one vowel would either have a faster reaction time when asked to identify that vowel or be more likely to identify an ambiguous stimulus as that vowel. I calculated these "frequency difference metrics" for all three continua.

For all participants, [a͡ɪ] had a higher token frequency than [ʌ͡i], which was in turn more fre-

	mean	sd	min	max
a͡ɪ+ɔ͡ɪ	7.77	0.004	7.760	7.779
a͡ɪ+ʌ͡i	7.53	0.012	7.503	7.546
ʌ͡i+ɔ͡ɪ	9.08	0.014	9.053	9.103

Table 3.3: Summary statistics for summed informativity for all three contrasts. All numbers are in bits.

quency than [ɔ͡ɪ]. Additionally, Table 3.2 gives information about the distribution of token frequency difference for the non-excluded participants in the study. It can be seen that the distributions for a͡ɪ-ɔ͡ɪ and a͡ɪ-ʌ͡i both are right-skewed and the distribution of ʌ͡i-ɔ͡ɪ is left-skewed.

3.2.2 Informativity

In contrast, while type informativity for [a͡ɪ] and [ʌ͡i] are negatively correlated, they are not exactly predicted by each other — $R^2 = .37$. Thus, as it is reasonable to posit that a higher summed informativity would lead to faster reaction times overall, I calculated these "summed informativity metrics" for all three continua.

For all participants, [ɔ͡ɪ] had the highest type informativity, followed by [ʌ͡i], then [a͡ɪ]. Additionally, Table 3.3 gives information about the distribution of summed informativity for the non-excluded participants in the study. It can be seen that all three distributions are left-skewed.

3.3 Eyetracking transformation

Unfortunately, despite increasing interest in eye tracking in multiple scientific disciplines, raw eye tracking data is still not plug-and-analyzable. I used EyeLink Data Viewer (SR Research Ltd., 2021) to manipulate the data into a usable format.

The raw eye tracking data was transformed via a Time Course (Binning) Analysis.[10] This collected looking data into bins of a set duration. I set the analysis to report, for each bin, for each image, how many samples had the participant looking at said image. I set the Bin Interval to 20ms, so that each bin would contain 10 samples,[11] and included all samples in fixations and saccades. I left the Maximum Number of Bins as 2000.

Because the experimental design allowed participants to self-initiate the playing of the stimulus, it was important to track when this happened for each trial and adjust for this offset. To do this, Danica had ExperimentBuilder output a `PLAY_SOUND` message to a timesynced log whenever a stimulus would begin to play, and Data Viewer was instructed to load these logs.[12] I then created a Message Report[13] that would include the `CURRENT_MESSAGE_TEXT` and `CURRENT_MESSAGE_TIME` for only `PLAY_SOUND` messages.

I then loaded the binned time course information and the adjustment data into R before correcting for the offset. This required me to first transform the ordered bins into time data once more before subtracting the offset and rebinning — if possible, I would recommend finding a way to avoid this step in future experiments. Ideally, the correction would have either been done in Data Viewer or entirely obviated by an alternative experimental flow. However, because the experiment was fairly lengthy, allowing participants to proceed at their own pace was preferable to forcing participants to follow an arbitrary pace and potentially collecting noisier data.

3.4 Chapter summary

In this chapter, I have presented the experimental design of the study, describing the participant recruitment process, the stimuli, the experimental flow for the training, calibration, and testing

[10] Analysis > Reports > Time Course (Binning) Analysis ...

[11] Recall the sampling rate was set to 500 Hz, or 2 ms per sample.

[12] Preference > Data Loading > Load ExperimentBuilder Log Messages

[13] Analysis > Reports > Message Report ...

blocks, and the post-experiment survey. I have also described how the survey data was transformed into lexical statistics for individual participants, as well as how the raw eye tracking data was aggregated for future analysis.

In the following chapters, I will discuss how the time course of decision making differed between the three continua, how the categorization response data differed between the three continua, and of course, the influence of the lexical predictors on time course and categorization.

CHAPTER 4

Eye tracking analysis and results

So, after all of this, what indeed does eye tracking tell us about marginal contrasts and the influence of the lexicon? In this chapter, I analyze 1.58 million bins of eye tracking observations (see §3.3 for details) collected from 31 participants. We will see that, surprisingly, participant lexicons do not much influence participant behavior.

4.1 Analysis of time course divergence

As a preliminary step, I wanted to understand how participants behaved on average with respect to categorization of endpoint stimuli. As the endpoint stimuli are the most phonetically dissimilar, differences are most noticeable when comparing endpoints.

Recall that participants started each trial by looking at a dot placed between the two images representing the nonce words. By understanding which image is being looked at and when, it is possible to see if participants are, on average, faster or slower at categorizing specific vowels.

Following Steffman (2020) and Steffman & Sundara (2024), in Fig. 4.1, I used the `ggplot2` package (Wickham, 2016) in R to plot the proportion looks to the left endpoint vowel for step 1 and step 10 of all three of the continua for all 31 participants. 95% confidence intervals are also shown. As is standard, step 1 is the left endpoint of the continuum, so we expect that participants will increase proportion looks to left over time for this continuum step. Step 10 is the right endpoint, so we expect that participants will decrease proportion looks to left over time for this step.

Additionally, I have indicated some landmarks which will be useful for future reference: the

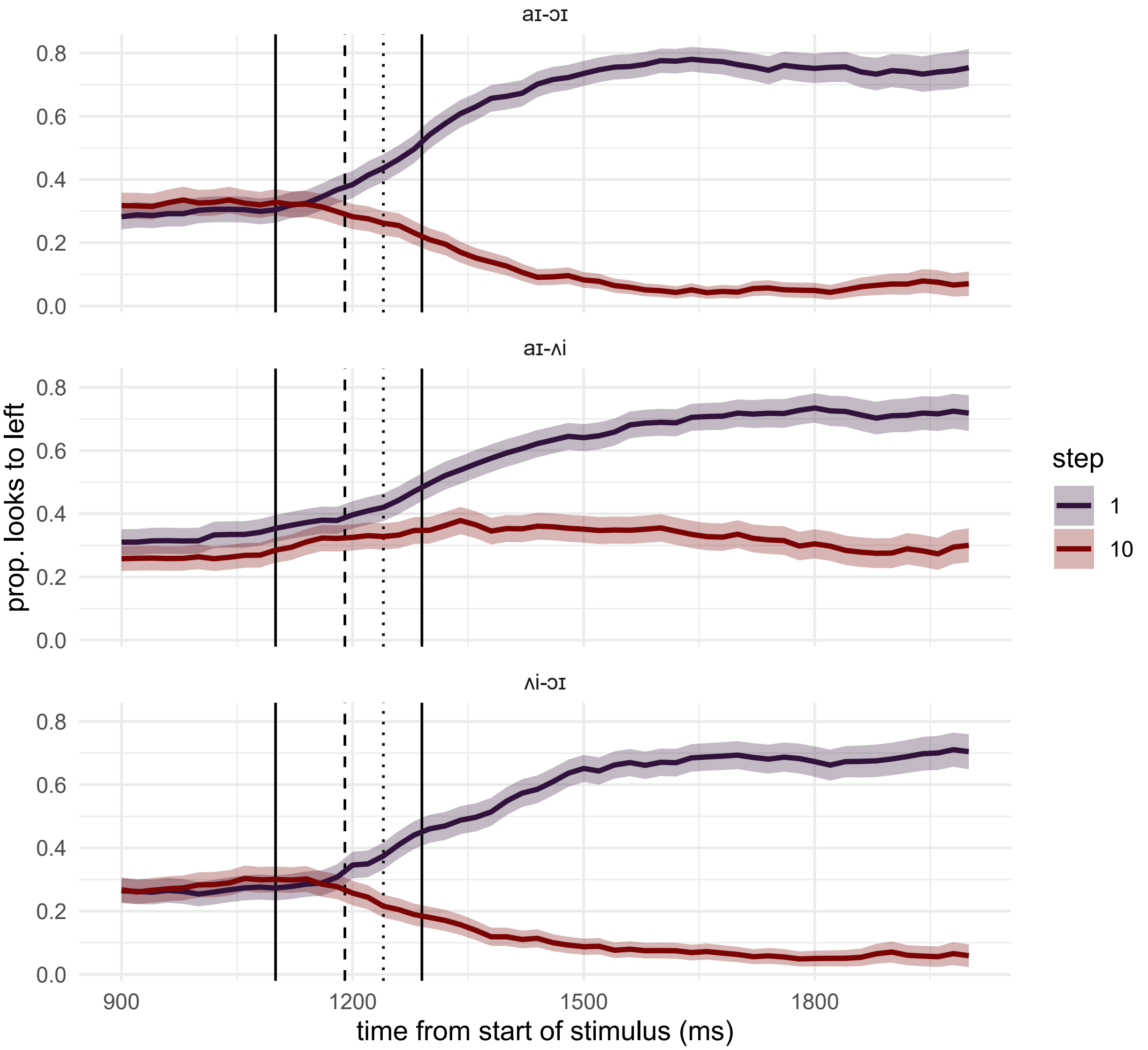


Figure 4.1: Proportion looks to the left vowel of the continuum, with 95% confidence intervals. The graphs begin at the onset of the vowel, and the solid lines represent a 200 ms offset for the duration of the vowel. The dashed line is the point of earliest divergence, which is within [a͡ɪ]-[ɔ͡ɪ], at 1190 ms. The dotted line is the point of latest divergence, which is within [a͡ɪ]-[ʌ͡i], at 1240 ms.

portion enclosed within the solid lines is a 200 ms offset of the duration of the vowel — as Allopenna et al. (1998) found that the reaction time to acoustic information is approximately 200 ms, this enclosed area represents the period of time wherein participants are likely reacting to the vowel. The dashed line is the earliest time coordinate where the 95% confidence intervals diverge for any of the three continua, and the dotted line is the latest such time coordinate.

Looking at Fig. 4.1, it can be immediately seen that the [a͡ɪ]-[ʌ͡i] continuum is not like the other two. First, the time of divergence for [a͡ɪ]-[ʌ͡i] is at 1240 ms, 50 ms later than the time of divergence for [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ], which is at 1190 ms.[1] This represents about a quarter of the duration of the vowel and indicates that participants needed to hear more of the vowel in this condition before they were able to begin categorization.

Further, it is apparent that proportion looks to left decreases over time for step 10 stimuli in both the [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ] continua but not in the [a͡ɪ]-[ʌ͡i] condition. Instead, proportion looks to left appears quite steady for step 10 in [a͡ɪ]-[ʌ͡i]. This behavior is different from predictions that would be obvious from the literature, and at first blush seems to indicate that participants may not be able to consistently distinguish between [a͡ɪ] and [ʌ͡i].

However, the middle [a͡ɪ]-[ʌ͡i] panel in Fig. 4.2 shows that looking behavior across the phonetic continuum does change over the steps, so it seems only that participants retain some amount of unconscious tendency to classify even fully naturalistic tokens of [ʌ͡i] as [a͡ɪ].

To give another perspective into participant willingness to classify tokens of [ʌ͡i] as [a͡ɪ], in Fig. 4.3, I have plotted proportion looks to left for all three continua on a per-step basis. These graphs show that a decrease in proportion looks to left appears as early as step 5 for [ʌ͡i]-[ɔ͡ɪ] (green line, middle left panel) and in step 6 for [a͡ɪ]-[ɔ͡ɪ] (gold line, middle right panel), but that no decrease is seen in any step for [a͡ɪ]-[ʌ͡i] (blue lines).

Finally, though Fig. 4.1 and Fig. 4.3 make it clear that participants are still looking at [a͡ɪ] even in step 10 of [a͡ɪ]-[ʌ͡i], these figures do not tell us if participants are also looking at [ʌ͡i] or if they are

[1] The exact point of divergence of [ʌ͡i]-[ɔ͡ɪ] is 1195 ms, but this 5 ms difference should not be considered reliable due to the rebinning process which was detailed in §3.3.

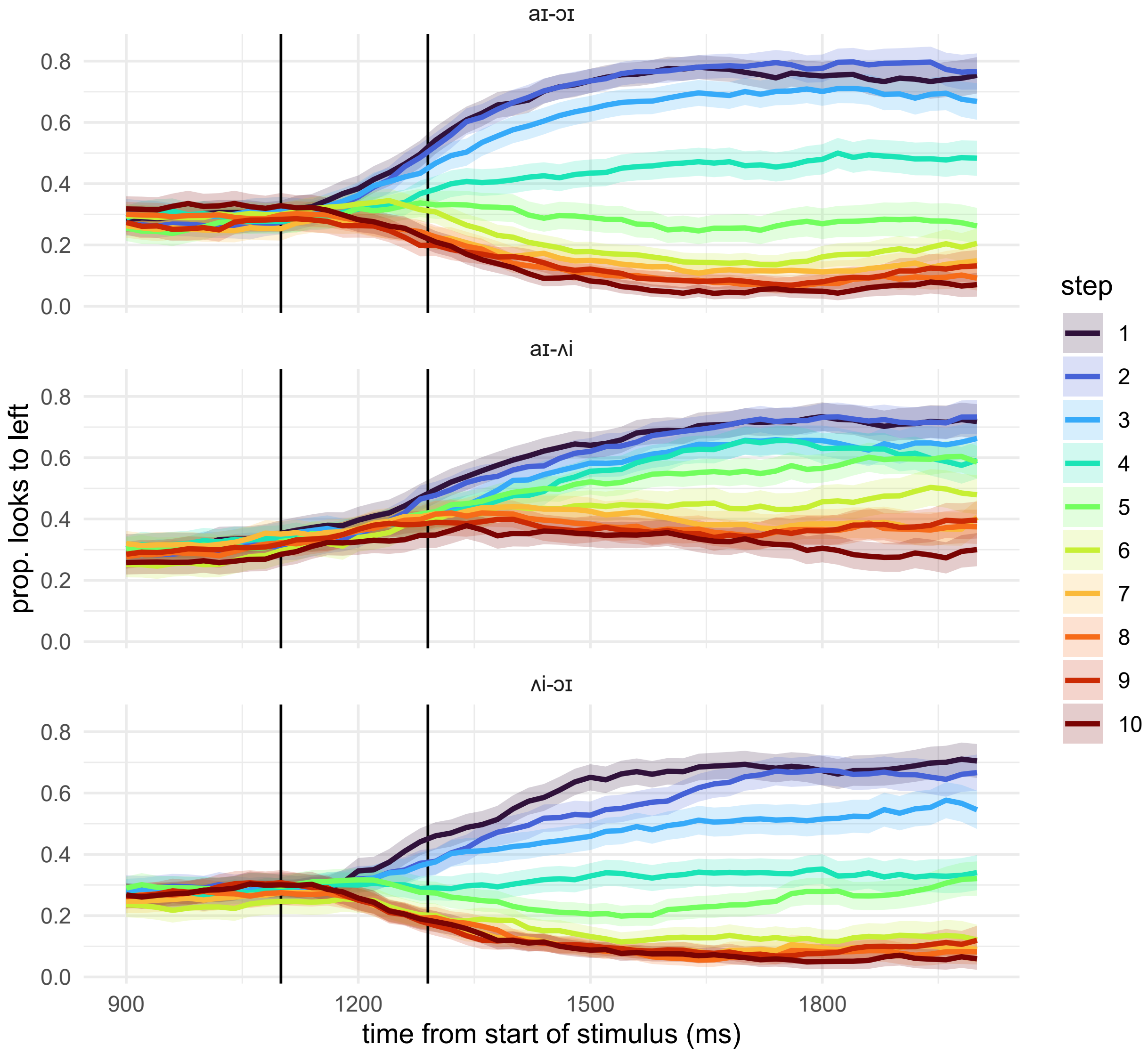


Figure 4.2: Proportion looks to the left vowel of the continuum, with 95% confidence intervals. The graphs begin at the onset of the vowel, and the solid lines represent a 200 ms offset for the duration of the vowel.

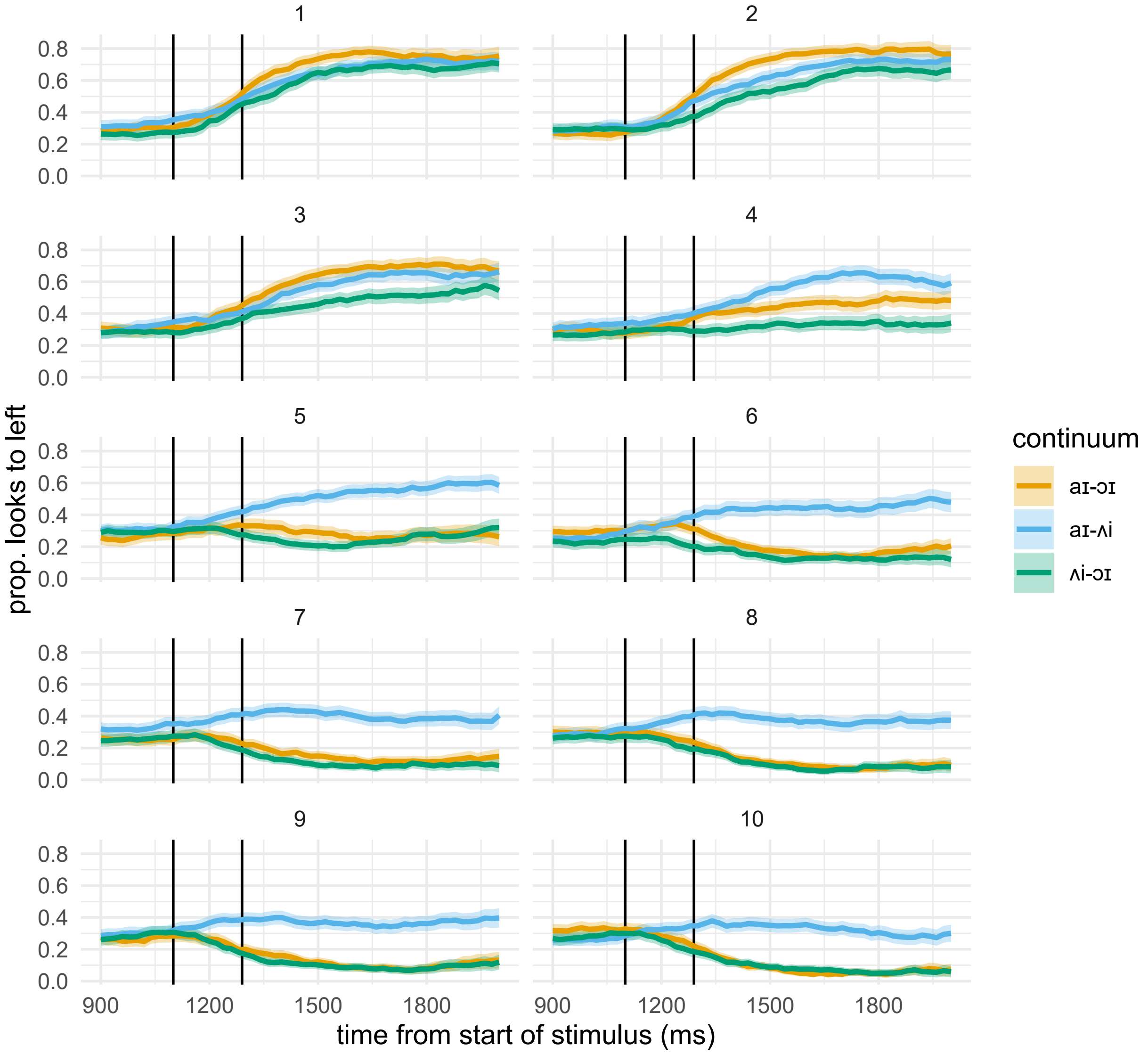


Figure 4.3: Proportion looks to the left vowel of the continuum, with 95% confidence intervals. The graphs begin at the onset of the vowel, and the solid lines represent a 200 ms offset for the duration of the vowel. Note how prop. looks to left begins to decrease over time in step 5 for [ʌ͡i]-[ɔ͡ɪ] and in step 6 for [a͡ɪ]-[ɔ͡ɪ].

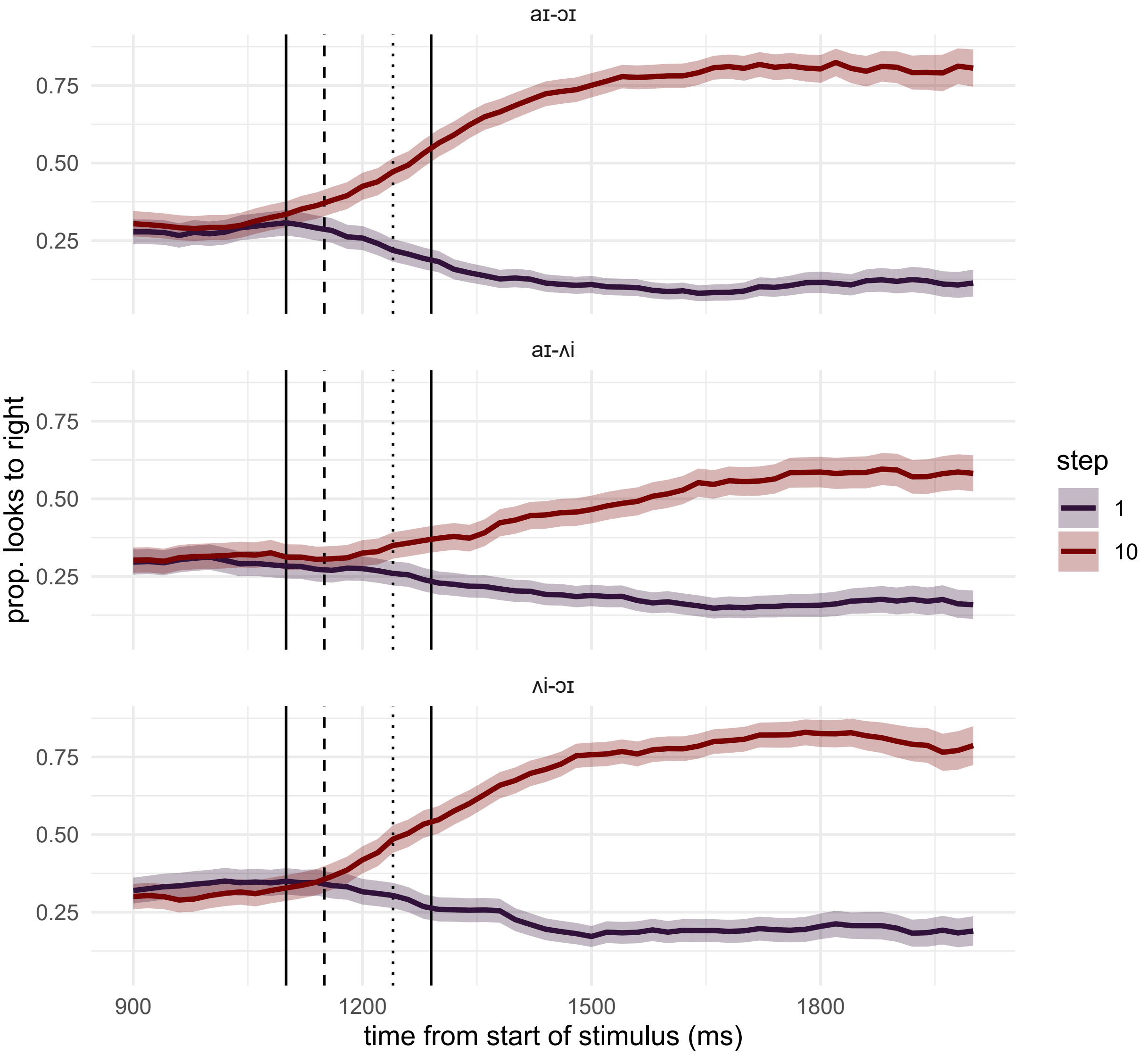


Figure 4.4: Proportion looks to the **right vowel** of the continuum, with 95% confidence intervals. The graphs begin at the onset of the vowel, and the solid lines represent a 200 ms offset for the duration of the vowel. The dashed line is the point of earliest divergence, which is within [a͡ɪ]-[ɔ͡ɪ], at 1150 ms. The dotted line is the point of latest divergence, which is within [a͡ɪ]-[ʌ͡i], at 1240 ms.

looking at nothing (i.e., if they are so confused they cannot determine what to look at). Thus, I plot in Fig. 4.4 proportion looks to **right**. In the middle panel of this figure, we can see that participants do increase in proportion looks to [ʌ͡i] for step 10 of that continuum, though they look to the right less consistently than they do for step 10 of the [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ] continua.

In sum, considering the time course of divergence shows that while participants treat the [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ] continua similarly, the [a͡ɪ]-[ʌ͡i] continuum reveals that, when the alternative is [a͡ɪ], [ʌ͡i] is frequently miscategorized. It is not that participants are unable to make a categorization judgment for [ʌ͡i]. Instead, it is misidentified as [a͡ɪ] at a high rate.

4.2 Analysis of lexical predictors

The effect of lexical variables on performance was assessed by model comparison of GAMMs (Generalized Additive Mixed Models). As a class of model, GAMMs are a type of GAM that incorporate random effects. The primary use case for GAMs and GAMMs is to model a response variable using nonlinear predictors, though GAMs are also capable of representing linear predictors. GAMs are capable of doing this because they model data using splines, series of summed functions that can approximate a wide variety of "wiggly"[2] functions — this is represented by the k and m parameters. k represents the number of basis points, so k must be greater to accurately represent more complex functions. m is the dimension of the penalty basis, and has the effect of penalizing squared derivatives of degree m of the spline.[3] Thus, when $m = 2$, the penalty is proportional to the squared second derivatives, preferring linear estimates; when $m = 1$, the penalty is proportional to the squared first derivatives, preferring point estimates.

In the limit, a GAM can accurately represent an arbitrarily complex predictor with high enough k and m, though the fitting function penalizes extreme values.[4]

[2] Note term of art.

[3] See Pedersen et al. (2019) for discussion on this point.

[4] Even so, the selection of k is important for the practical reason that high k or m results in more complex calculations

In particular, GAMs are appropriate for modeling data with high autocorrelation — data where observations are highly correlated with others in close temporal or spatial proximity. Autocorrelation is ubiquitous when observing physical processes such as eye tracking. Because the eye must move in real space, what a participant looks at within one 20 ms bin will be spatially proximal to what they look at in the next. As such, GAMs have been frequently used in linguistics to model eye tracking behavior.

For the models here, the dependent variable is a normalized preference measure. This approach is taken following Reinisch & Sjerps (2013) and Steffman (2020), who employed similar visual world paradigms. This measure, which I will call "bias to left" or just "bias", is the empirical logit transformation from Barr (2008). As Barr (2008) says, this transformation usefully "filters out eye-movement based dependencies in the data", but requires aggregation over multiple observations. Since the data being modeled has already been binned, we can use the extant bins as spans of observations to aggregate over. Thus, for each bin, the dependent variable $\eta\prime$ is

$$\eta\prime = ln(\frac{y + .5}{n - y + .5})$$

where y is the number of samples within the bin where a participant is looking at the left item and n is the number of samples within the bin that the participant is looking at any item. As the binning procedure resulted in 10 samples/bin, this means that $\eta\prime$ has a range of ± 3.04, with positive values interpreted as bias towards the left item and negative items interpreted as bias towards the right.

For each vowel continuum, I used the `mgcv` package (Wood, 2017) in R to fit data from all 31 participants to 2 GAMMs, a "baseline" model, plus a "lexical" model containing the lexical predictors. Fig. 4.5 contains representative model schemata. The baseline GAMMs model bias as a tensor product smooth between time and step and include two random effects for participant

during model fitting, so setting appropriate values may greatly reduce coffee consumption while waiting for results. There is a `k.check()` function to check the appropriateness of k but it was not always able to arrive at an optimal setting for each predictor, as the function is nondeterministic.

```
continuum_baseline_gam = bam(bias ~
 te(t_ms, step) +
 s(participant, bs="re", m=1) +
 s(trial, participant, bs="fs", m=1, k=5),
data = data_continuum, method = "ML")

continuum_lexical_gam = bam(bias ~
 te(t_ms, step) +
 s(continuum_tok_f, m=2) + s(continuum_typ_i, m=2) +
 ti(t_ms, step, continuum_tok_f, k=3, m=2) +
 ti(t_ms, step, continuum_typ_i, k=3, m=2) +
 s(participant, bs="re", m=1) +
 s(trial, participant, bs="fs", m=1, k=5),
data = data_continuum, method = "ML")
```

Figure 4.5: Model schemata for baseline models and lexical models. Note that `bs="re"` indicates a random linear effect and `bs="fs"` indicates a random smooth effect. Discussion of `k` and `m` are provided in text.

continuum	% deviance explained (base)	% deviance explained (lex)	ΔAIC
[a͡ɪ]-[ɔ͡ɪ]	16.9	17	312
[a͡ɪ]-[ʌ͡i]	14.4	14.6	728
[ʌ͡i]-[ɔ͡ɪ]	6.36	6.58	1103

Table 4.1: Model comparison results between the baseline and lexical models. Here, ΔAIC is $\text{AIC}_{\text{base}} - \text{AIC}_{\text{lex}}$, so all lexical models have lower AIC and are thus preferred for each continuum. However, the % deviance explained does not differ much between the models.

and trial within participant. The tensor product smooth is shorthand for two main effect smooths and their interaction, so `te(t_ms, step)` is analogous to `t_ms*step` in a lmer model, and the random intercepts, modeled as smooths with $m = 1$, are analogous to `(1|participant)` and `(trial|participant)`. In short, the baseline GAMMs predict bias as an interaction between time and continuum step, with random intercepts for participant and trial within participant.

The lexical GAMMs further include main effect smooths for the lexical predictors (token frequency difference and summed type informativity) and tensor product interactions between time, step, and the lexical predictors. A tensor product interaction is a manually specified interaction with no implied main effects, so `ti(t_ms, step, continuum_tok_f)` is equivalent to specifying `t_ms:step:continuum_tok_f` in an lmer model. All terms including lexical predictors are modeled as smooths with $m = 2$, as the lexical effects that are theoretically predicted are monotonic. To further enforce conformity with theoretical predictions, the `ti()` terms have $k = 3$ as well. Thus, the lexical GAMMs have additional main effects for frequency difference and summed informativity — for the sake of interpretability – as well as 3-way interactions between those main effects, time, and continuum step.

I used the `compareML()` function from the `itsadug` package (Van Rij et al., 2022) to conduct model comparison. For all three continua, I found that the lexical model outperformed the baseline model with $p \approx 0$. Within each lexical model, the lexical main effects were not statistically significant while the interactions were. However, the amount of improvement seen in the lexical models was rather minimal. Indeed, as seen in Table 4.1, the % deviance explained, which indicates goodness of fit to the data, only increases between 0.1–0.3% with the addition of the lexical predictors.

Visualization of data and predictions is an invaluable part of any analysis, so I also used the `pvisgam()` function from `itsadug` to plot the partial effect of `t_ms` and `step` on bias — that is, the predicted bias across `t_ms` and `step`, holding other predictors fixed. These contour plots are provided in Fig. 4.6 and it can be seen that many facts about these predictions are congruent with the time course divergence plots presented in Fig. 4.1. Though here I show only the plots from the lexical models, the corresponding plots from the baseline models look very similar (as should be expected).

These plots are modeled after maps and have `t_ms` on the x-axis, `step` on the y-axis, and bias on the z-axis, as color. As with maps, contour lines are also drawn on the graph to easily understand the shapes on the plot.

I find these graphs to be a succinct and intuitive way of displaying predictions for bias, as they contain compact representations of many of the insights that can be gleaned from time course divergence data. For example, from Fig. 4.6, it is clear that the predicted bias ranges from between 4 and -2 in the middle [a͡ɪ]-[ʌ͡i] panel and that the predicted bias is more restricted for [a͡ɪ]-[ʌ͡i] than for other continua. Indeed, the bias is positively skewed, indicating that participants are predicted to be generally biased towards [a͡ɪ], even when the stimulus is a naturalistic [ʌ͡i] — this is an alternate view of what was seen in the [a͡ɪ]-[ʌ͡i] middle panel in Fig. 4.1, as proportion looks to left did not decrease in step 10 the way it did in other panels.

We see that the contour lines are very close together at t>1700 ms between steps 3 and 6 in the top [a͡ɪ]-[ɔ͡ɪ] panel. This indicates that, within this time frame, predicted bias is quite different across these steps and this was also seen in the [a͡ɪ]-[ɔ͡ɪ] top panel in Fig. 4.1, where we see that the "steady state" values of proportion looks to left are most dissimilar between steps 3 and 6.

We can also see that the overall shape of the [ʌ͡i]-[ɔ͡ɪ] bottom panel in Fig. 4.6 seems down-

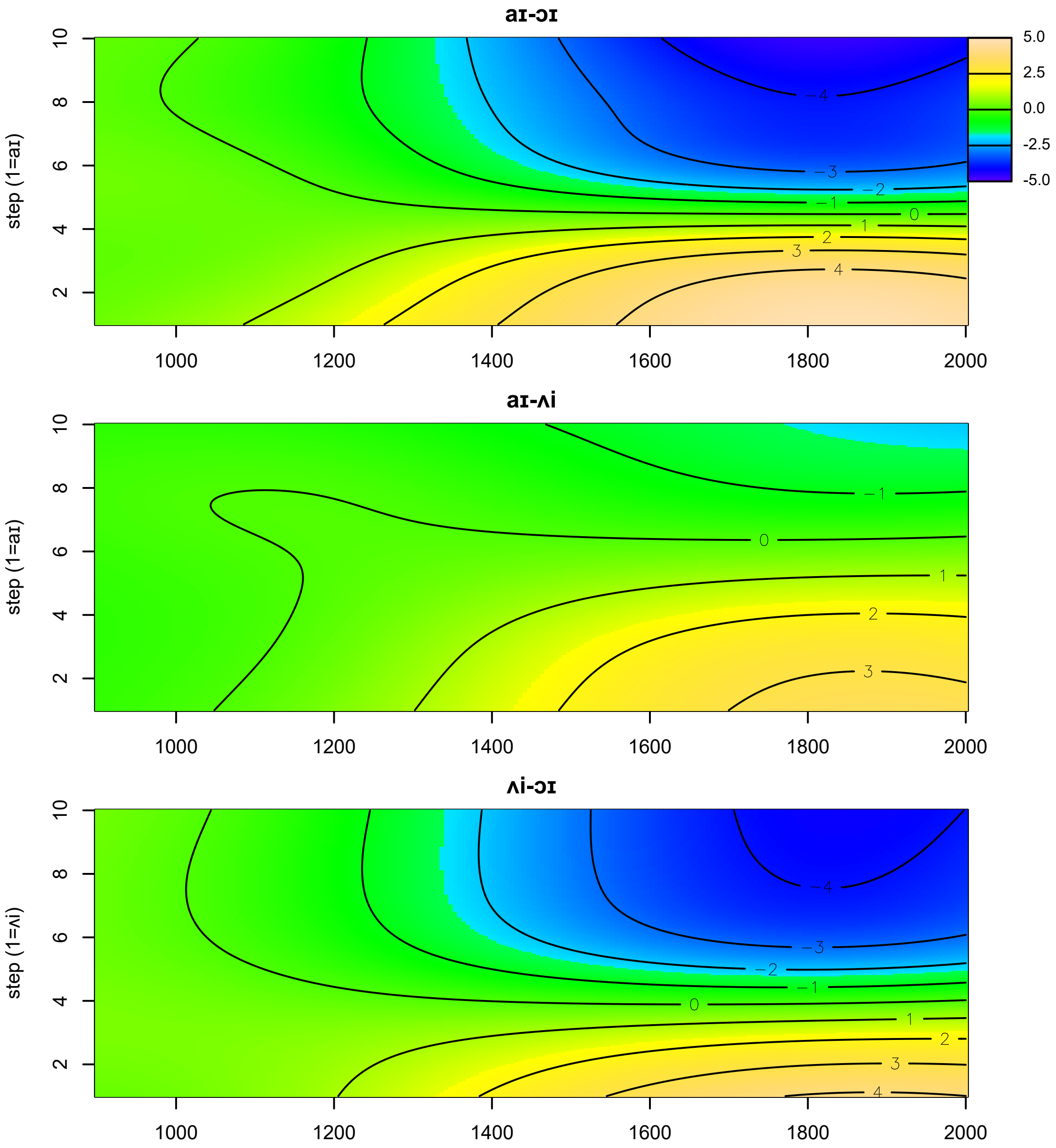


Figure 4.6: Partial effects plots for `bias∼te(t_ms, step)` from the lexical models. Time (ms) is on the x-axis, step is on the y-axis, and color indicates bias. Steps range from the left member of the continuum at step 1 to the right member at step 10.

shifted relative to that of the [a͡ɪ]-[ɔ͡ɪ] top panel. That is, it seems that the positive bias peak is further off the bottom of the map in the bottom panel. This tracks with the fact that proportion looks to left was lower for step 1 in the [ʌ͡i]-[ɔ͡ɪ] bottom panel of Fig. 4.1 relative to step 1 in the [a͡ɪ]-[ɔ͡ɪ] top panel.

Finally, it is possible to imagine what these contour plots would look like if circumstances were different. Considering a scenario where participants were unable to distinguish between [a͡ɪ] and [ʌ͡i], participants would likely be biased towards [a͡ɪ] responses. The resulting contour plot might then look like a version of the [a͡ɪ]-[ʌ͡i] panel where the yellow extends all the way up the y-axis, or nearly so. If participants were completely confounded by the task and had no bias, the panel would be green throughout.

These partial effects plots are generally congruent with theoretically based predictions. The top [a͡ɪ]-[ɔ͡ɪ] panel looks like categorical perception: there is no real bias until the participant starts to react to the stimulus at around 1100 ms with bias rapidly changing across steps in the middle of the continuum and more gradually changing towards the edges. The middle [a͡ɪ]-[ʌ͡i] panel is clearly different with the bias lines much more even spaced across the continuum, slower decision-making at endpoint steps, and less extreme values of bias overall. The bottom [ʌ͡i]-[ɔ͡ɪ] panel is somewhere in between.

4.2.1 Visualizations of lexical interactions

Having established that the interpretation of contour plots is not fiendishly obtuse, I turn now to the visualization of the partial effects plots for the lexical variables. Alas, the visualization of these partial effects makes it clear that my selected lexical predictors do not affect eye tracking in a straightforwardly interpretable way.

Recall that I posited in §3.2.2 that higher summed informativity would lead to heightened perceptual sensitivity, thus faster reaction times overall. If this were true, then we can imagine a partial effects graph for high summed informativity: in this plot, we would expect to see an effect begin-

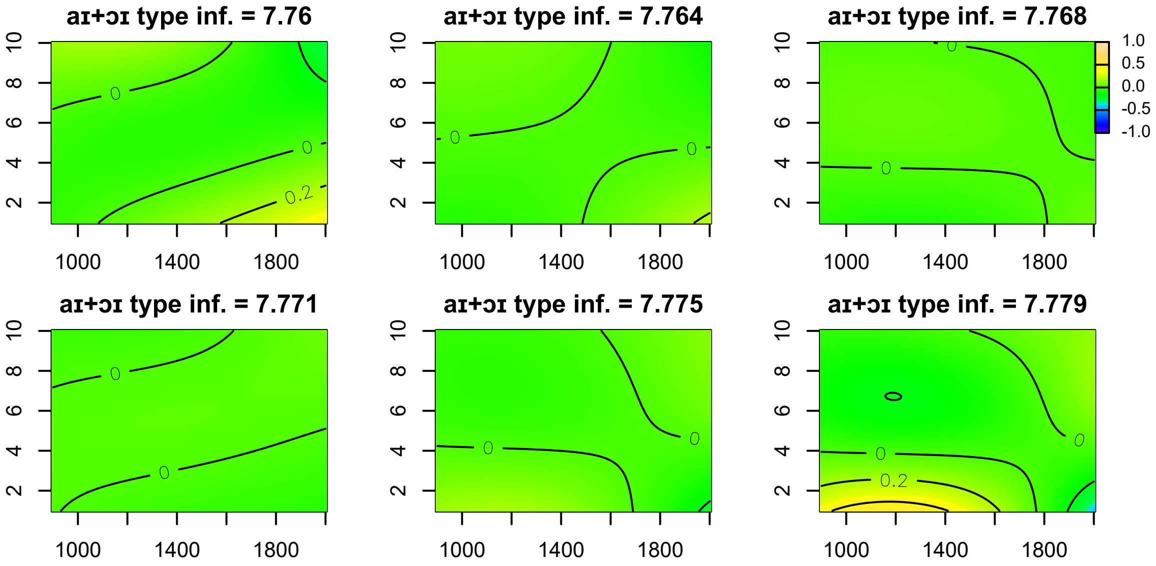


Figure 4.7: Partial effects plots for [a͡ɪ]-[ɔ͡ɪ] informativity, plotted at 6 equally spaced points representing the range of observed variation of this variable. Contour lines are at every 0.2.

ning no earlier than 1100 ms. The effect would not affect intermediate continuum steps, would only affect steps that already see eventual non-zero bias, and would serve to overall push contour lines leftwards or earlier in time.

In Figs. 4.7–4.9, I have plotted the partial effect of summed informativity for the three continua at 6 equally spaced values representing the range of observed variation of this variable.[5] Examining them even cursorily, it is clear that we do not see any such pattern for any continuum and that the three partial effects are each unlike the others.

Indeed, within the 900–2000 ms interval we have been looking at, there is not much to interpret. The upper left panel of Fig. 4.7 ([a͡ɪ]-[ɔ͡ɪ]) shows that participants with low summed informativity for [a͡ɪ] and [ɔ͡ɪ] have a slightly increased bias towards [a͡ɪ] in [a͡ɪ]-like steps but that this effect only holds after 1600 ms. Moreover, this bias disappears for people with intermediate summed informativity before returning in the participants with highest summed informativity, but only between 900-1600 ms. This does not follow from any theory I am aware of.

[5] That is, the minimum and maximum observed values and 20%, 40%, 60%, and 80% between them.

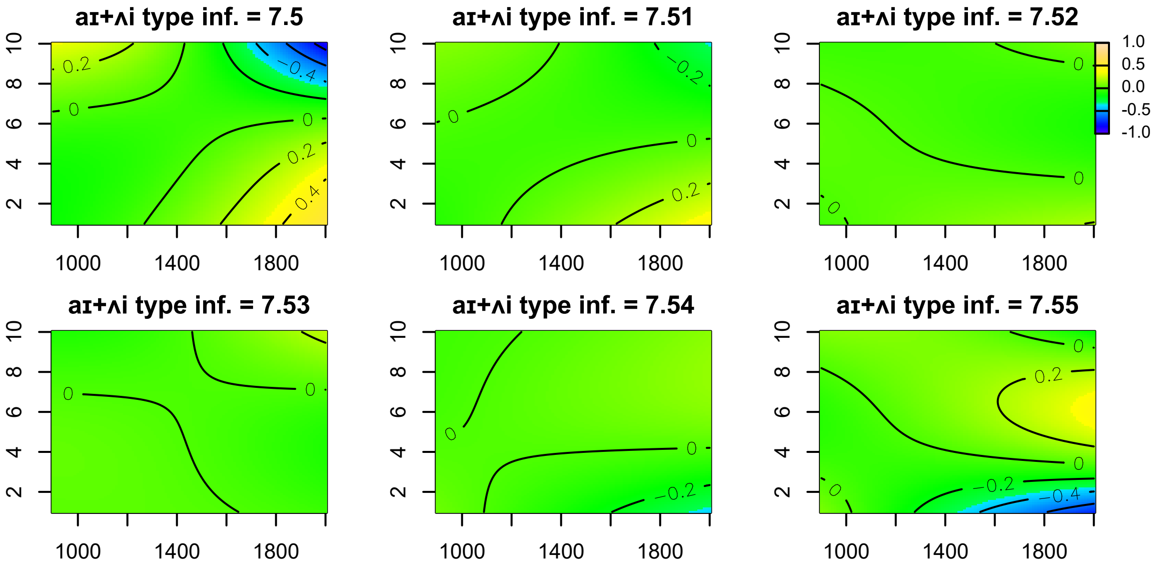


Figure 4.8: Partial effects plots for [a͡ɪ]-[ʌ͡i] informativity, plotted at 6 equally spaced points representing the range of observed variation of this variable. Contour lines are at every 0.2.

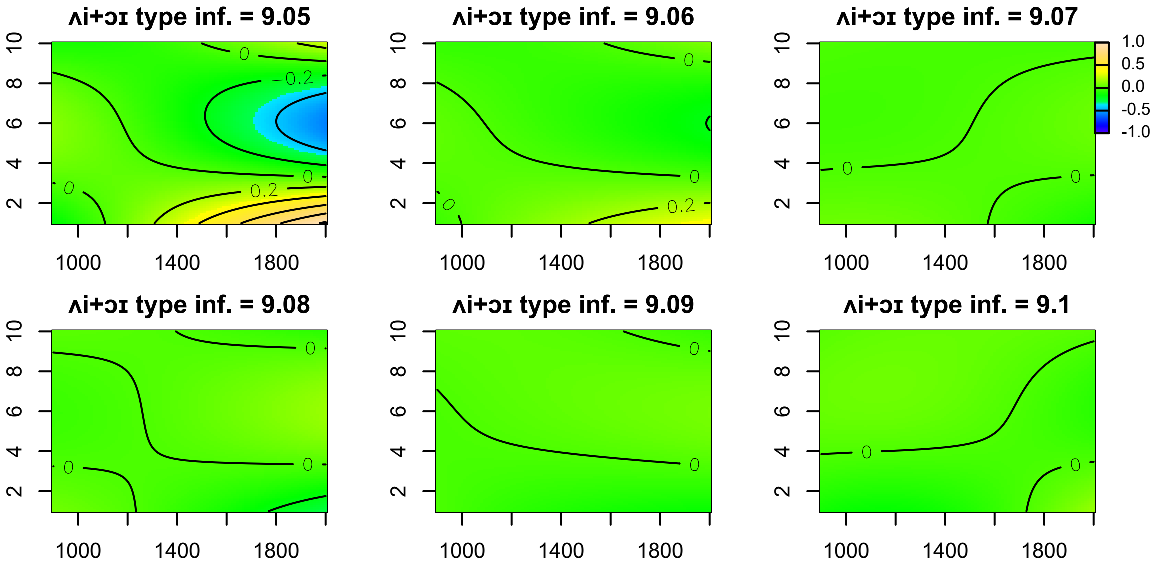


Figure 4.9: Partial effects plots for [ʌ͡i]-[ɔ͡ɪ] informativity, plotted at 6 equally spaced points representing the range of observed variation of this variable. Contour lines are at every 0.2.

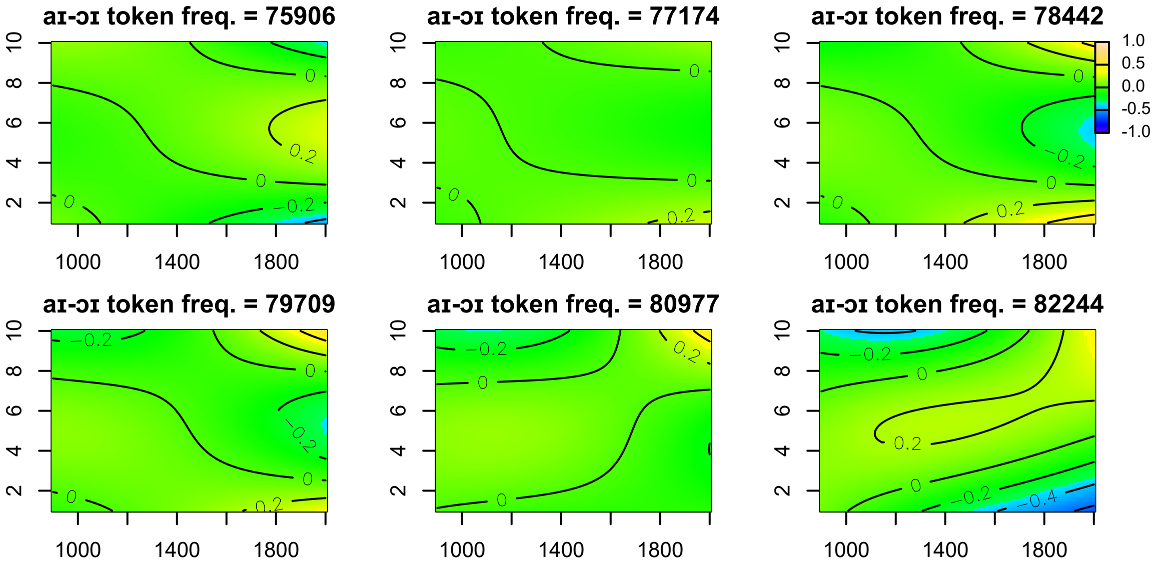


Figure 4.10: Partial effects plots for [a͡ɪ]-[ɔ͡ɪ] frequency, plotted at 6 equally spaced points representing the range of observed variation of this variable. Contour lines are at every 0.2.

Similar issues of interpretability plague each of the summed informativity partial effects. The lower right panel of Fig. 4.8 ([a͡ɪ]-[ʌ͡i]) shows that bias towards [a͡ɪ] weakens in step 1 for participants with high summed informativity for [a͡ɪ] and [ʌ͡i], which does not make sense if higher summed informativity corresponds to stronger phonemic categories. The lower panels of Fig. 4.9 ([ʌ͡i]-[ɔ͡ɪ]) shows that there is no effect for high summed informativity for [ʌ͡i] and [ɔ͡ɪ], which certainly does not follow from the previous supposition.

It so happens that the situation does not improve for the frequency difference partial effects. Recall that I posited in §3.2.1 that a participant with a frequency distribution more skewed towards one vowel would either have a faster reaction time when asked to identify that vowel or be more likely to identify an ambiguous stimulus as that vowel. The first option would translate into something similar to what was predicted for summed informativity, but crucially the reaction time difference would only affect opposite vowels at opposite ends of the frequency spectrum; the second would result in a blanket effect from 1100ms onwards, though the direction of the effect would again be opposite at opposite ends of the spectrum.

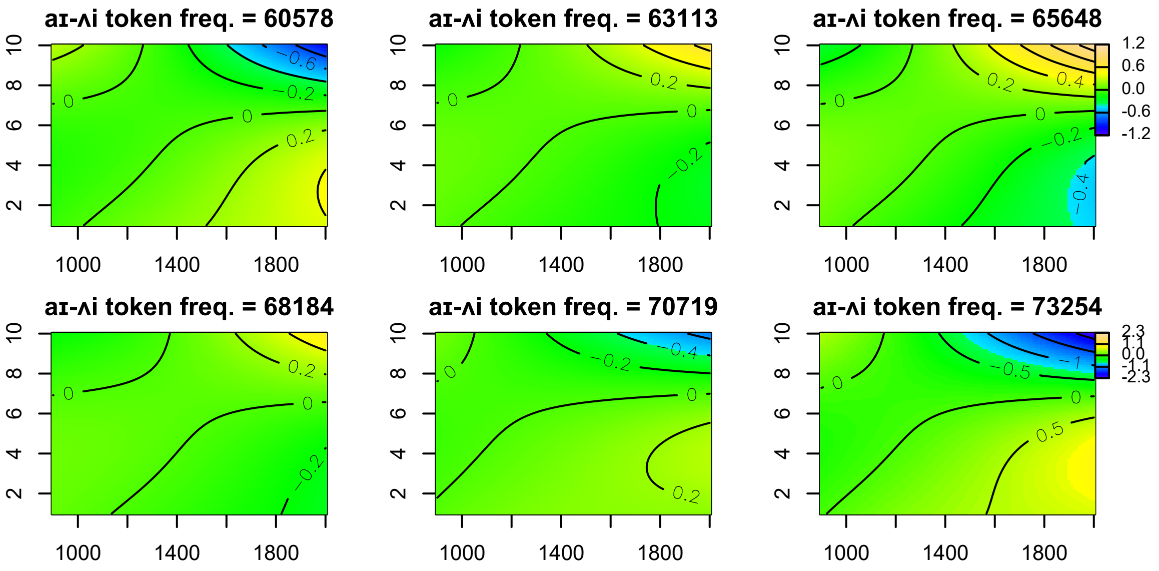


Figure 4.11: Partial effects plots for [a͡ɪ]-[ʌ͡i] frequency, plotted at 6 equally spaced points representing the range of observed variation of this variable. Contour lines are at every 0.2 but for last panel, where at every 0.5.

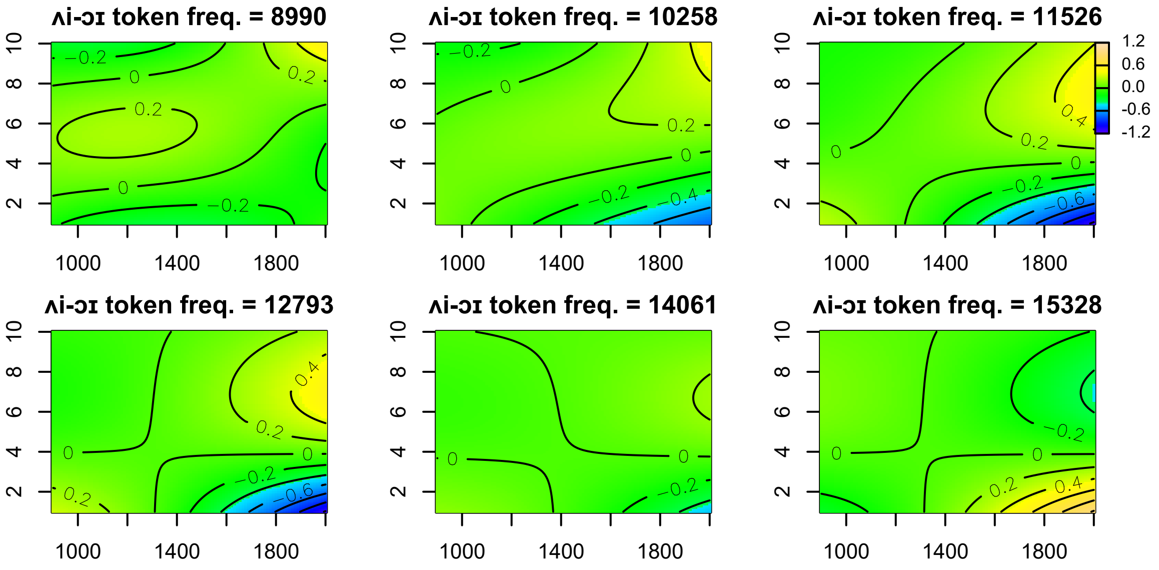


Figure 4.12: Partial effects plots for [ʌ͡i]-[ɔ͡ɪ] frequency, plotted at 6 equally spaced points representing the range of observed variation of this variable. Contour lines are at every 0.2.

However, looking at Figs. 4.10–4.12, it is again quite obvious that not only are neither of the predicted patterns seen, the interpretability issues are perhaps worse. For we can see that there is flip-flopping in each of the frequency difference partial effects plots — in Fig. 4.10 ([a͡ɪ]-[ɔ͡ɪ]) , bias towards [a͡ɪ] in step 1 is negative for participants with low [a͡ɪ] token frequency, but then positive for participants with intermediate frequencies, then negative again for those with the highest token frequencies.

Not only are there flip-flops in Fig. 4.11 and Fig. 4.12, but the results are also contradictory. In Fig. 4.11, it is the difference between token frequencies for [a͡ɪ] and [ʌ͡i] that is plotted, so a lower number indicates more frequent [ʌ͡i] and less frequent [a͡ɪ]. Thus, bias towards [ʌ͡i] in step 10 is negative for participants with high [ʌ͡i] token frequency, but then positive for participants with lower frequencies, then negative again for those with the lowest [ʌ͡i] token frequencies.

In Fig. 4.12, it is the difference between token frequencies for [ʌ͡i] and [ɔ͡ɪ] that is plotted, so a lower number indicates *less frequent* [ʌ͡i] and *more frequent* [ɔ͡ɪ] In Fig. 4.12, bias towards [ʌ͡i] in step 1 is zero for participants with low [ʌ͡i] token frequency, but then negative for participants with higher frequencies, then positive for those with the highest [ʌ͡i] token frequencies.

4.3 Chapter summary

Looking at the time course divergence data, it is clear that the participants treat the three continua differently, with the [a͡ɪ]-[ɔ͡ɪ] continuum most closely hewing to expectations for categorical perception and the [a͡ɪ]-[ʌ͡i] continuum most divergent from said expectations. We see that participants are generally slower to make decisions about [ʌ͡i]-[ɔ͡ɪ] stimuli — approximately 50 ms slower, which corresponds to approximately 25% of the vowel. We also see that participants are reluctant to fully commit to [ʌ͡i] categorizations when attending to [a͡ɪ]-[ʌ͡i] stimuli. These same findings are also present in the contour plots in Fig. 4.6, giving us good confidence that the overall experimental design was not flawed.

However, when we look at the status of the lexical predictors, we run into some problems.

First, though model comparison indicated that the lexical variables are all statistically significant predictors of bias, the inclusion of the lexical predictors improved model fit little in absolute terms. Attempting to make sense of this by visualizing the lexical partial effects, we find that the lack of interpretability of these plots contrasts strongly with the commonsensical interpretation so easily arrived at for the main effects contour plots. This difficulty in interpretation and lack of evidence congruent with any of the predictions made for these predictors suggests that the selected variables are not truly meaningfully affecting how participants are behaving, regardless of the fact that statistical methods indicated significance for all of them.

CHAPTER 5

Categorization analysis and results

In the immediately preceding chapter, I presented the results of the eye tracking analysis. Though we saw patterns in the divergence data and contour plots which showed that participants are treating the three continua differently from each other, we also found that the lexical statistics that I calculated did not have an interpretable effect on the eye tracking data.

Thus, I turn now to the question: do informativity and frequency have an interpretable effect on any of the collected data? In this chapter, I will present the categorization curve results from the eye tracking experiment, an analysis of 14697 categorization decisions collected from 31 participants. We will find that although natural groups of participants emerge from the categorization results, these groups do not naturally fall out of the lexical statistics on hand, calling into question the appropriateness of summed informativity and frequency difference as predictors for phonemic strength.

5.1 Analysis of categorization data

As we had also collected the categorization responses during the eyetracking study, I constructed a dataset representing only categorization responses by filtering the eyetracking results such that each trial was represented only once, giving me a total of 14697 decisions across 31 participants.

Using the `ggplot2` package, I plotted this data as shown in Fig. 5.1. A common design language is used for all visualizations of response curves in this dissertation. The continua are color coded, with [a͡ɪ]-[ɔ͡ɪ] in gold, [a͡ɪ]-[ʌ͡i] in blue, and [ʌ͡i]-[ɔ͡ɪ] in green. The x-axis is the continuum

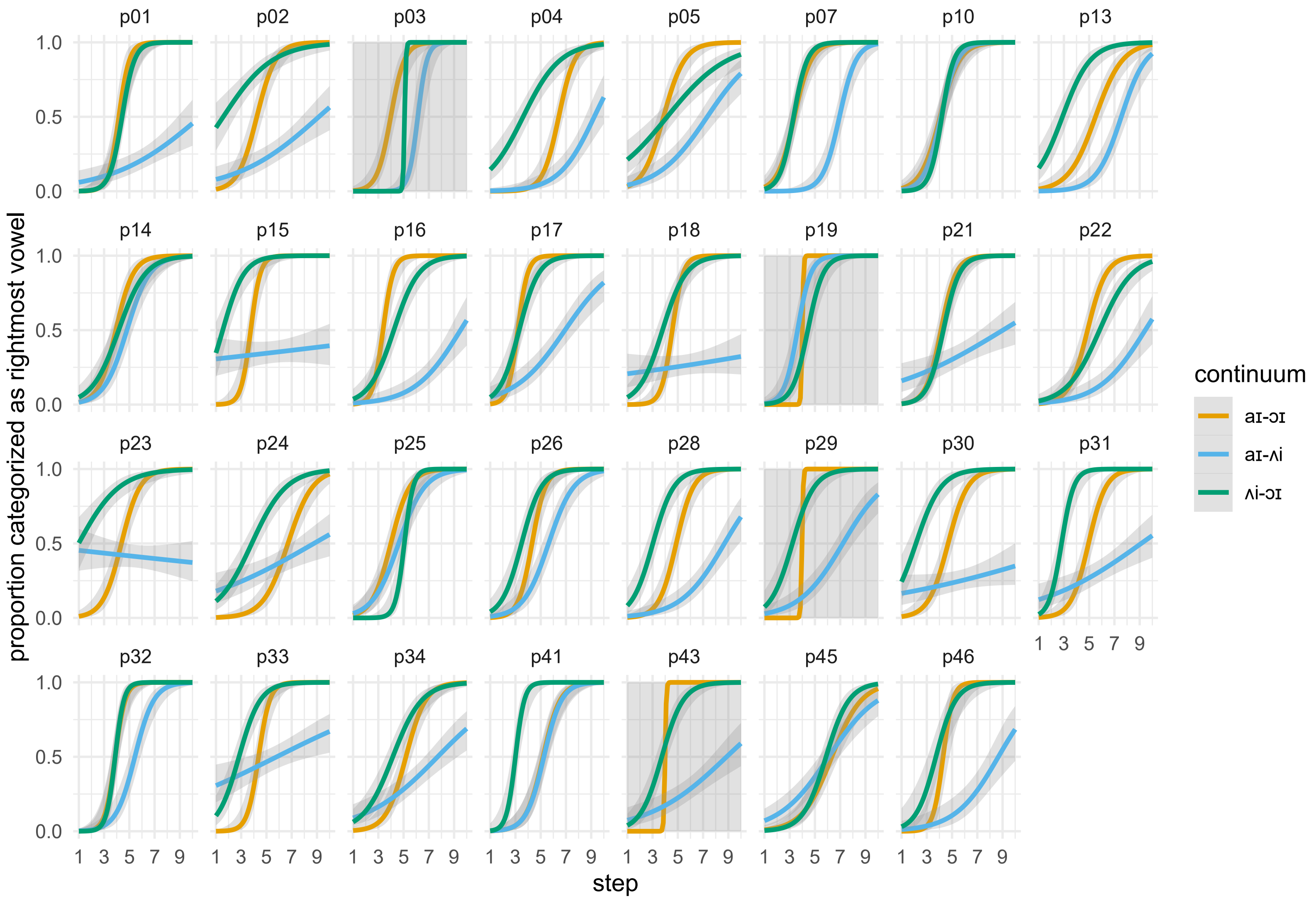


Figure 5.1: Response curves for all three continua for each participant. Continuum step is on the x-axis, with step 1 being the left vowel endpoint, and proportion categorization as left vowel is on the y-axis. 95% CIs are included.

step, with step 1 being the left vowel endpoint and step 10 being the right vowel endpoint. Thus, step 1 for the green, [ʌ͡i]-[ɔ͡ɪ] continuum is a naturalistic [ʌ͡i] and step 10 for the blue, [a͡ɪ]-[ʌ͡i] continuum is the same [ʌ͡i]. The y-axis is the proportion of trials within each step categorized as the right/step 10 vowel — so for the gold, [a͡ɪ]-[ɔ͡ɪ] continuum, the y-axis represents the proportion of tokens classified as [ɔ͡ɪ]. Thus, the general expectation is that steps closer to the left vowel endpoint will have fewer categorizations as the right vowel and y-values will rise from left to right. All participants are shown in Fig. 5.1, with response curves and 95% confidence intervals (CI) for each continuum plotted separately for each.[1]

Looking at Fig. 5.1, we can immediately see that although overall classification is as expected, with almost all y-values increasing as we progress from left to right, individual participants behave quite differently from each other. These differences in behavior are fairly complex and, when comparing all three continua, few participants behave identically to other participants. As such, the remainder of this section will go through some pairwise comparisons between conditions before discussing overall takeaways.

5.2 [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ]

We will first contrast behavior for [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ], as this comparison directly shows how substitution of a marginal phoneme, [ʌ͡i], for a strong phoneme, [a͡ɪ], affects categorization. For reference, I present in Fig. 5.2 only the response curves for these two continua — Fig. 5.2 is the same as Fig. 5.1 but without the blue, [a͡ɪ]-[ʌ͡i] lines.

Recall that in §4.1, we noted that the [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ] continua seem to behave similarly

[1] Note that some participants show what is called "complete separation" in their categorization of certain continua. For example, participant 3 has complete separation in their green [ʌ͡i]-[ɔ͡ɪ] response curve and participants 19, 29, and 43 have complete separation in their gold [a͡ɪ]-[ɔ͡ɪ] response curves. In this data, complete separation is when there is a point along the x-axis before which all responses are 0 and after which all responses are 1. This results in a situation where no confidence interval can be meaningfully computed, so `ggplot2` instead gives a confidence interval that spans [0,1] for all x-values. Unfortunately, though this makes the response curves for these participants harder to read, I am unable to suppress this behavior.

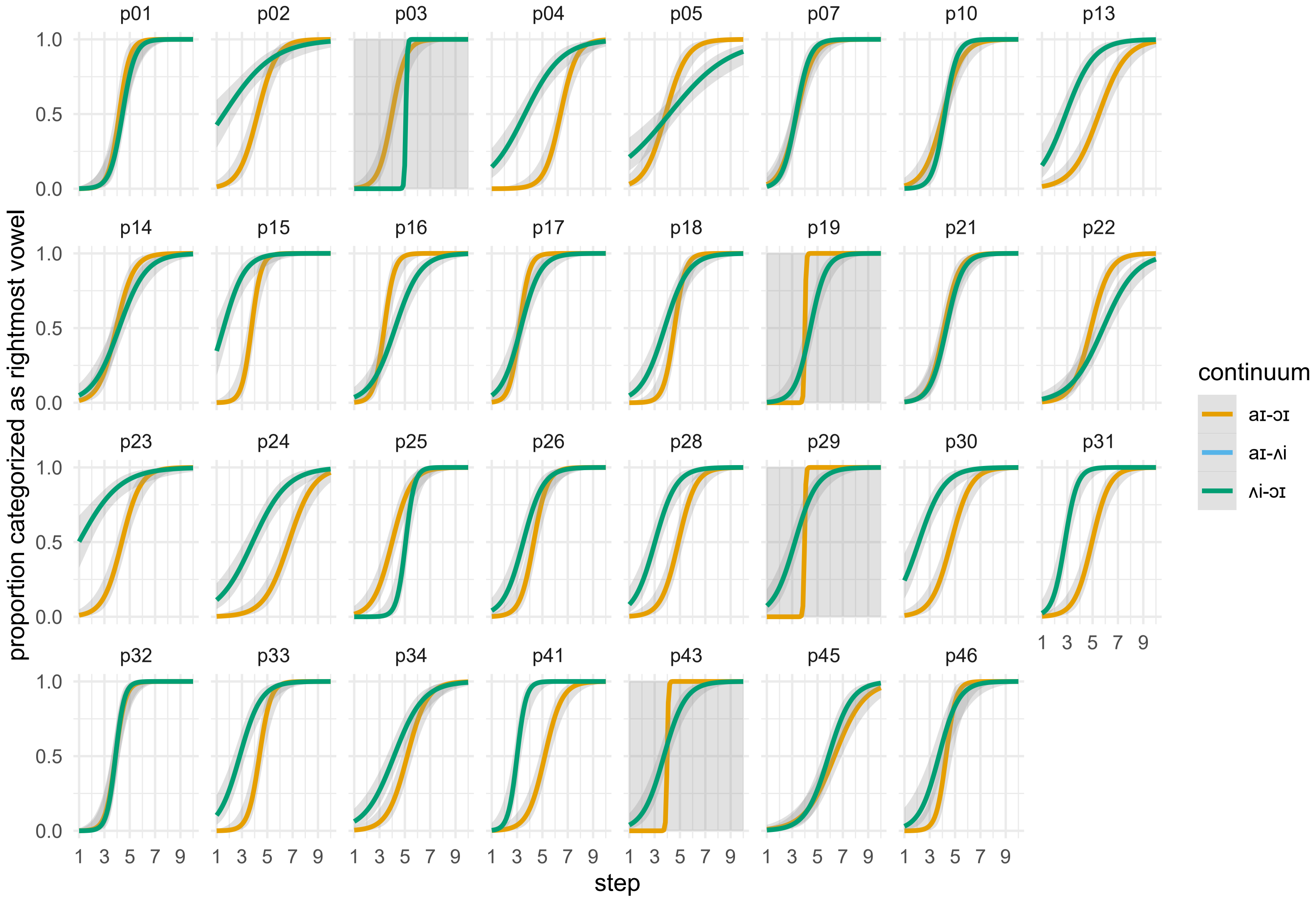


Figure 5.2: [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ] response curves for all participants.

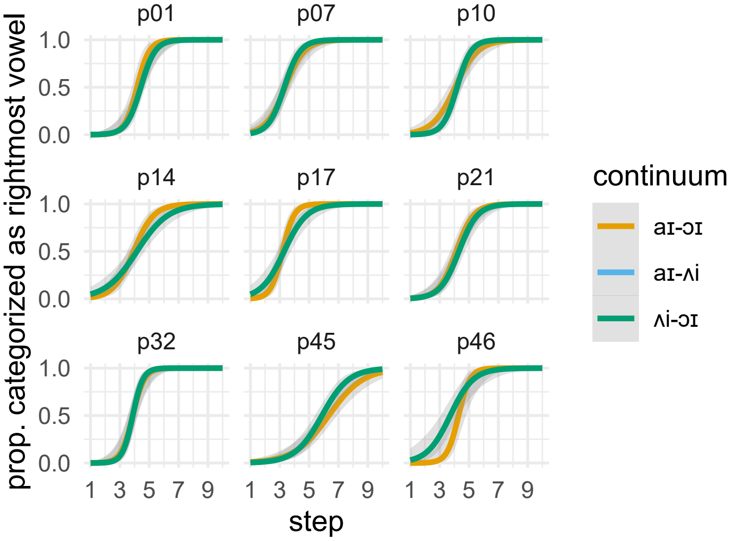


Figure 5.3: Response curves for all participants whose CI for [a͡ɪ]-[ɔ͡ɪ] overlaps with their CI for [ʌ͡i]-[ɔ͡ɪ] for all x values.

with respect to the divergence data. Looking at Fig. 5.2, however, we find that some participants treated stimuli from these continua quite differently. I will discuss three groups of interest: those with complete overlap between their [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ] response curves, those reluctant to commit to [ʌ͡i] categorizations, and the one with a plausibly linear categorization curve.

5.2.1 Complete overlap between [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ]

It appeared from Fig. 4.1 that participants treated step 1 of these continua near-identically, and we can see from Fig. 5.2 that indeed, there are participants whose gold [a͡ɪ]-[ɔ͡ɪ] and green [ʌ͡i]-[ɔ͡ɪ] categorization curves are not statistically distinguishable. Plotted in Fig. 5.3, these participants have CIs that overlap at all points across the two continua and number 9/31 or 29%.

One might conjecture that these participants may be hearing [ʌ͡i]-like stimuli and treating them as if they were [a͡ɪ]-like, and this could be sensible if participants with complete overlap had [a͡ɪ]-[ʌ͡i] lexical frequency distributions skewed towards [a͡ɪ]. Yet, participants with complete overlap have lower [a͡ɪ]-[ʌ͡i] token frequency difference measures than those without complete overlap (two-sample t test $t(23.6) = -2.09, p = .046$) and thus also have higher [ʌ͡i] token frequency ($t(23.6) = 2.09, p = .046$) and lower [a͡ɪ] token frequency ($t(23.6) = -2.09, p = .046$).[2]

Participants with complete overlap had slightly lower summed [a͡ɪ]+[ʌ͡i] type informativity ($t(11.8) = -2.62, p = 0.022$), though there was no statistically significant difference in [a͡ɪ] type informativity ($t(10.4) = -0.13, p = 0.902$) or [ʌ͡i] type informativity ($t(12.1) = 2.06, p = 0.061$) between the groups. This finding is not contrary to expectations, as lower summed [a͡ɪ]+[ʌ͡i] type informativity should result in worse contrast between [a͡ɪ] and [ʌ͡i], but we will see that it is the only finding that tracks with expectations.

5.2.2 Low commitment to [ʌ͡i]

In §2.2.1, I discussed how a high a or low d represent unwillingness to posit some category for extreme stimuli and raised the possibility that this would be observed for categorization of marginal phonemes. And lo, though some participants treated step 1 of these continua near-identically, we can see from Fig. 5.2 that a number of participants have green lines that do not rest at 0 at step 1. In other words, they were reluctant to classify the [ʌ͡i] endpoint as [ʌ͡i].

However, what is a reasonable boundary for reluctance? It must be acknowledged that participants do not always have 100% accuracy in endpoint categorization, even in cases where categorical perception is strongly expected, such as [a͡ɪ]-[ɔ͡ɪ]. Thus, to be confident in saying that a participant was reluctant to classify the [ʌ͡i] endpoint as [ʌ͡i], we should compare participant accuracy against the least accurate [a͡ɪ]-[ɔ͡ɪ] endpoint classification. In this experiment, that happens to be particpant 45, whose 95% CI for step 10 of the [a͡ɪ]-[ɔ͡ɪ] continuum ranges from 0.889 to 0.985.

Thus, I categorize participants whose CI for step 1 is wholly greater than 0.015 and those whose CI for step 10 is wholly lesser than 0.985 as “reluctant to commit”.

For step 1 of [ʌ͡i]-[ɔ͡ɪ], these participants number 16/31 or 52%. These participants are shown

[2] Recall that all [ʌ͡i] tokens in the corpus are derived from original [a͡ɪ] tokens, so there is a perfect, inverse correlation between the token frequency difference metric and [ʌ͡i] token frequency and a perfect, positive correlation with [a͡ɪ] token frequency.

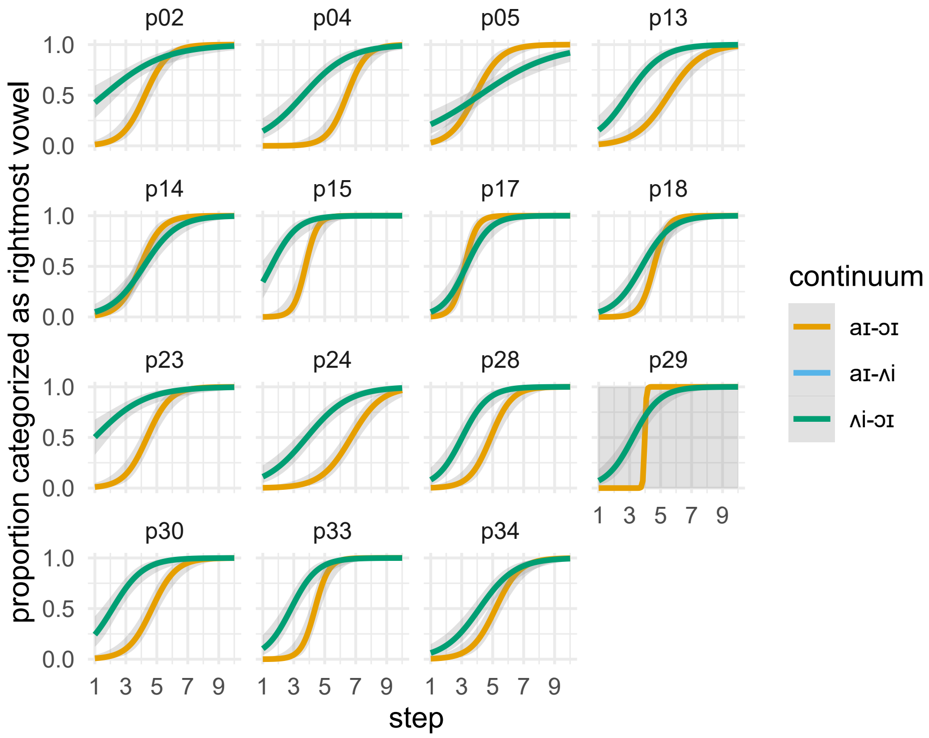


Figure 5.4: Response curves for all participants whose CI for step 1 responses for [ʌ͡i]-[ɔ͡ɪ] is wholly greater than 0.015.

in Fig. 5.4, and compared to all other participants, these individuals have no statistically significant difference in [a͡ɪ]-[ʌ͡i] token frequency difference measures ($t(27.6) = 0.59$, $p = .559$) and thus no difference in [ʌ͡i] token frequency or [a͡ɪ] token frequency. Thus, my experiment does not find evidence to support the hypothesis that lexical frequency has an effect on a.

Compared to participants who were willing to commit to [ʌ͡i], participants with low commitment to [ʌ͡i] had no statistically significant difference in summed [a͡ɪ]+[ʌ͡i] type informativity ($t(30.0) = -0.42$, $p = .678$), [a͡ɪ] type informativity ($t(30.0) = 0.25$, $p = .803$) or [ʌ͡i] type informativity ($t(30.0) = 0.44$, $p = .665$).

It is interesting to see that though some participants have a CI that is close to the cutoff of 0.015, three have a CI containing 0.5. That is, participants 2, 15, and 23 were so hesitant to commit to an [ʌ͡i] classification that they performed at chance even when listening to a naturalistic [ʌ͡i].

5.2.3 Phonetic categorization in [ʌ͡i]-[ɔ͡ɪ]

Finally, it can be seen in Fig. 5.2 that one individual, participant 5, is unique in having an effectively linear response "curve" for their green [ʌ͡i]-[ɔ͡ɪ] categorizations — a sort of response classically associated with categorization of phones instead of phonemes. In §2.2.1, I said that k, the steepness of the response curve, should be low for a participant if the predictability of [ʌ͡i] in their lexicon is high. However, as there is only one such person, it is impossible to conduct a significance test on any of the lexical variables and the hypothesis that links k to informativity cannot be argued for or against with this datum alone.[3]

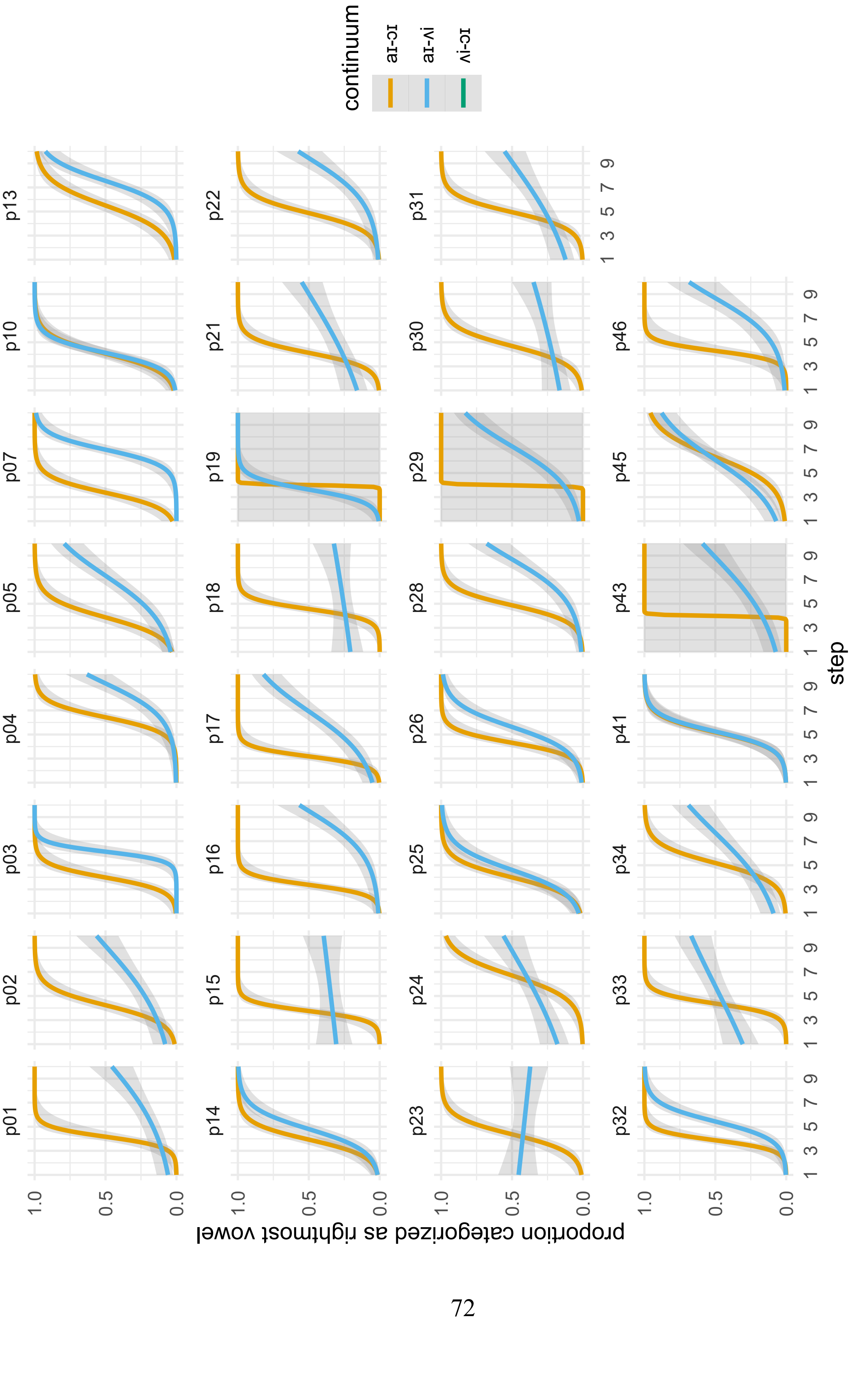


Figure 5.5: [a͡ɪ]-[ɔ͡ɪ] and [a͡ɪ]-[ʌ͡i] response curves for all participants.

5.3 [a͡ɪ]-[ɔ͡ɪ] and [a͡ɪ]-[ʌ͡i]

We will now contrast behavior for [a͡ɪ]-[ɔ͡ɪ] and [a͡ɪ]-[ʌ͡i]. Again, for reference, I present in Fig. 5.5 only the response curves for these two continua. Fig. 5.5 is the same as Fig. 5.1 but without the green, [ʌ͡i]-[ɔ͡ɪ] lines.

Here, I will first discuss the nine participants with evidence of categorical perception of [a͡ɪ]-[ʌ͡i]. Following that, I will discuss three telescoping groups of participants of interest — those with low commitment to [ʌ͡i], those with a plausibly linear response curve in [a͡ɪ]-[ʌ͡i], and those with a plausibly *flat* response curve in [a͡ɪ]-[ʌ͡i] — before discussing the participants with a nonlinear but also non-sigmoidal response curve.

5.3.1 Categorical perception in the [a͡ɪ]-[ʌ͡i] continuum

Looking at Fig. 5.5, only nine individuals, participants 3, 7, 10, 14, 19, 25, 26, 32, and 41 had blue, [a͡ɪ]-[ʌ͡i] response curves that reach a proportion of .985 or higher in step 10 of the [a͡ɪ]-[ʌ͡i] continuum and appear to be fully sigmoidal — having both a lower and upper asymptote. These 29% of the participants are presented in Fig. 5.6.

When compared as a group to participants who never achieve a high commitment to [ʌ͡i] using the definition of "reluctant to commit" given in §5.2.2, these individuals have no statistically significant difference in [a͡ɪ]-[ʌ͡i] token frequency difference measures ($t(26.7) = -1.49$, $p = .147$) and thus no difference in [ʌ͡i] token frequency or [a͡ɪ] token frequency. There was also no statistically significant difference in summed [a͡ɪ]+[ʌ͡i] type informativity ($t(20.8) = 0.41$, $p = .684$), [a͡ɪ] type informativity ($t(23.7) = 1.41$, $p = .172$) or [ʌ͡i] type informativity ($t(21.8) = -0.12$, $p = .906$). Thus, my experiment does not find evidence to support the hypothesis that lexical informativity has an effect on k.

[3] This participant's behavior comes as a surprise to me, as I would not have expected to see this behavior in the [ʌ͡i]-[ɔ͡ɪ] continuum but rather in the [a͡ɪ]-[ʌ͡i] continuum. Indeed, in §5.3.2 I discuss how we do see this behavior in abundance in [a͡ɪ]-[ʌ͡i]. Perhaps participant 5 is, though lacking a fully phonologized [ʌ͡i], in possession of a prodigious ear.

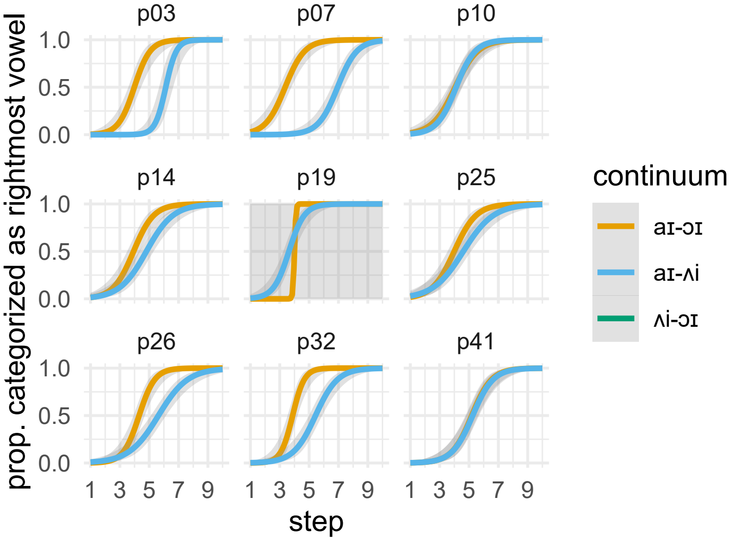


Figure 5.6: Response sigmoids for all participants whose CI upper bound for step 10 responses for [a͡ɪ]-[ʌ͡i] is greater than 0.985.

Perhaps this is the result of internal heterogeneity in individual behavior. Comparing how these participants categorized [a͡ɪ]-[ʌ͡i] stimuli (blue) and [a͡ɪ]-[ɔ͡ɪ] stimuli (gold), we can see that for some participants the categorical perception of ambiguous stimuli as [a͡ɪ] spans more steps in the [a͡ɪ]-[ʌ͡i] continuum than in the [a͡ɪ]-[ɔ͡ɪ] continuum. This is most evident for participant 7, who categorically classifies steps 1–4 in [a͡ɪ]-[ʌ͡i] as [a͡ɪ] and steps 7-10 in [a͡ɪ]-[ɔ͡ɪ] as [ɔ͡ɪ], but only categorically classifies the opposite endpoints (step 10 in [a͡ɪ]-[ʌ͡i] as [ʌ͡i], step 1 in [a͡ɪ]-[ɔ͡ɪ] as [a͡ɪ]).

This behavior seems to indicate the presence of some Ganong-like effect. The classic Ganong effect is the tendency to perceive ambiguous signals as the option most likely to result in a real word (Ganong, 1980). However, Steffman & Sundara (2024) find that when a listener is presented with two nonce words anchoring a phonetic continuum, they shift their categorization curve to favor the nonce word with higher (biphone) probability. Thus, I suggest we can recast Steffman & Sundara's (2024) result as evidence for an Extended Ganong effect, the tendency to perceive ambiguous signals as the more *wordlike* option.

Though it has been argued that the classic Ganong effect should be delayed for phonemic decision tasks, Kingston et al. (2016) found that a Ganong effect can be observed as early as 350 ms after stimulus onset in a categorization task similar to the one conducted here, and Steffman & Sundara (2024) found that biphone probability affected eye movements “as early as 400–500 ms after [the point of stimulus difference]”. However, while the unigram and bigram probabilities in Table. 3.1 show that [klɔ͡ɪɾɚ], the [ɔ͡ɪ] nonce word, is a more probable word than [kla͡ɪɾɚ], the [a͡ɪ] nonce word, the same is interestingly not true for [a͡ɪ] and [ʌ͡i]. Indeed, the probability of [klʌ͡iɾɚ] is the highest among the three nonce words, suggesting that we ought to see a greater number of steps about the [ʌ͡i] endpoints being classified as [ʌ͡i]. While this prediction holds for some of these nine individuals, it is intriguing that even participants with solid evidence for categorical perception in [a͡ɪ]-[ʌ͡i] do not seem to all treat [ʌ͡i] as they do [ɔ͡ɪ] — that is, their [ʌ͡i]s are not affected by even the Extended Ganong effect.

Unfortunately, as there are so few participants that can be compared against each other, it is not possible to meaningfully statistically compare those with an [ʌ͡i] affected by the Extended Ganong effect with those unaffected.

5.3.2 Low commitment to [ʌ͡i]

On the flip side, looking at Fig. 5.5, it is evident that a majority of the participants have blue, [a͡ɪ]-[ʌ͡i] response curves that never reach 1, even at step 10. In a situation parallel to that which was already discussed in §5.2.2, these participants were reluctant to classify the [ʌ͡i] endpoint as [ʌ͡i] in the [a͡ɪ]-[ʌ͡i] continuum, though here we see a low d instead of a high a due to the direction of the continuum.

Indeed, using the same definition of “reluctant to commit” from §5.2.2, 22/31 or 71% of participants had CIs for step 10 of [a͡ɪ]-[ʌ͡i] that were wholly less than .985. Moreover, of these 22 individuals, 12 had a CI for step 10 that includes 0.5, meaning 39% of all participants were so hesitant to commit to an [ʌ͡i] classification that they performed at chance even when listening to a naturalistic [ʌ͡i].

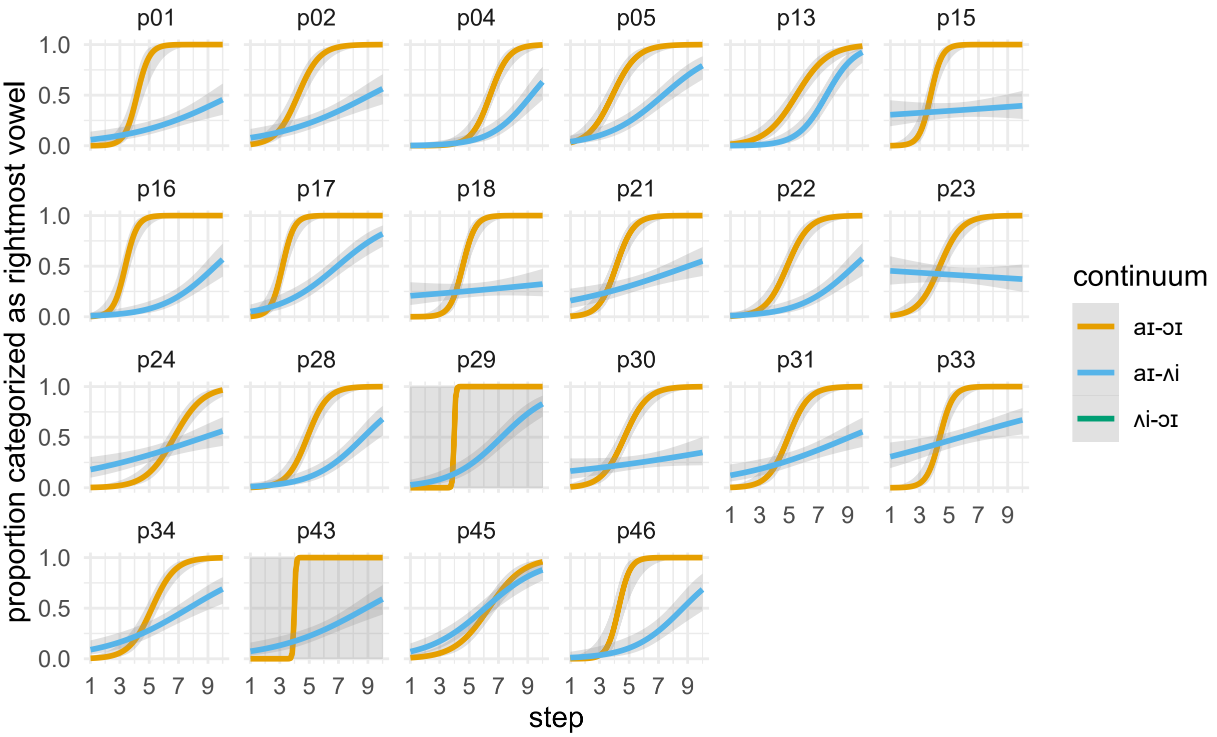


Figure 5.7: Response curves for all participants whose CI for step 10 responses for [a͡ɪ]-[ʌ͡i] is wholly less than 0.985.

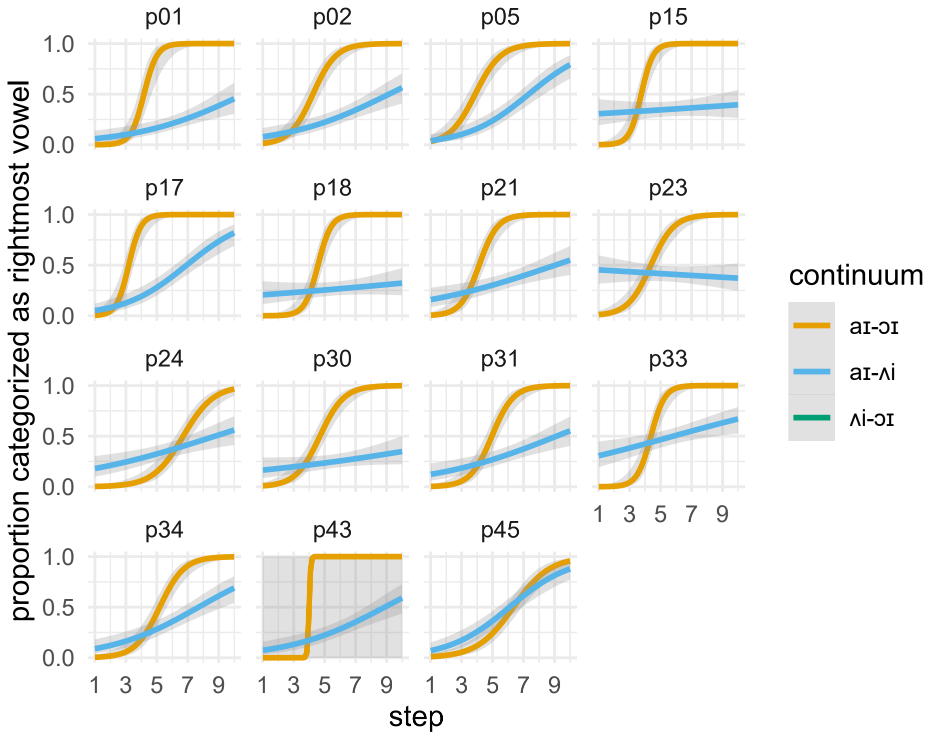


Figure 5.8: Response curves for all participants whose response curve [a͡ɪ]-[ʌ͡i] is plausibly linear, including those whose curve is plausibly flat.

These participants are shown in Fig. 5.7, and we can see that their [a͡ɪ]-[ʌ͡i] response curves are quite heterogeneous, so it will be challenging to draw any conclusions on the basis of low commitment to [ʌ͡i] alone. Of course, as these participants are the complement to the participants discussed in the previous section, statistically comparing these participants to those who achieved higher commitment to [ʌ͡i] results in the same numbers in the immediately preceding section, just with opposite signs on the t-statistic. Thus, my experiment does not find evidence to support the hypothesis that lexical frequency has an effect on d.

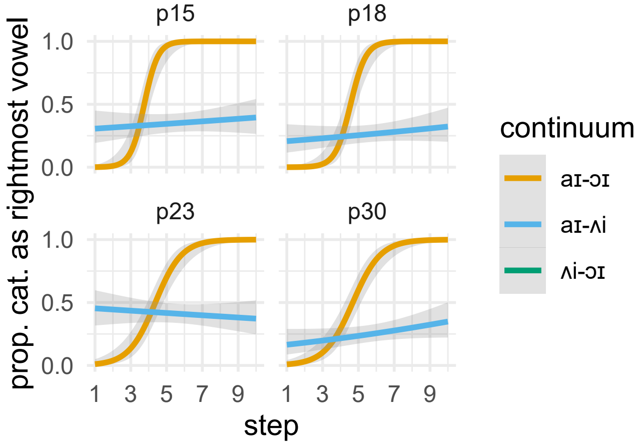


Figure 5.9: Response curves for all participants whose response curve [a͡ɪ]-[ʌ͡i] is plausibly flat.

5.3.3 Linear and flat responses to [a͡ɪ]-[ʌ͡i]

However, that said, there are two subgroups here that merit further attention. First, the 15 individuals highlighted in Fig. 5.8 have blue, [a͡ɪ]-[ʌ͡i] response curves that have not only low k parameters, but are also plausibly linear. By this, I mean that it is possible to draw a single straight line across the x-axis that fits entirely within the 95% confidence interval. Linear response curves are typically understood to be the result of gradient phonetic categorization, but compared to participants with nonlinear response curves, these 48% of all participants have no statistically significant difference in [a͡ɪ]-[ʌ͡i] token frequency difference measures ($t(18.5) = -1.13$, $p = .272$) and thus no difference in [ʌ͡i] token frequency or [a͡ɪ] token frequency. There was also no statistically significant difference in summed [a͡ɪ]+[ʌ͡i] type informativity ($t(26.1) = -0.50$, $p = .621$), [a͡ɪ] type informativity ($t(30.0) = -0.70$, $p = .487$) or [ʌ͡i] type informativity ($t(25.5) = -0.17$, $p = .863$). Thus, my experiment does not find evidence to support the hypothesis that lexical informativity has an effect on k.

There are also four individuals, highlighted in Fig. 5.9 that have blue, [a͡ɪ]-[ʌ͡i] response curves that look flat. By this, I mean that it is possible to identify a single y-value which is within the 95% confidence interval for all x-values. These participants do not seem to have been able to reliably

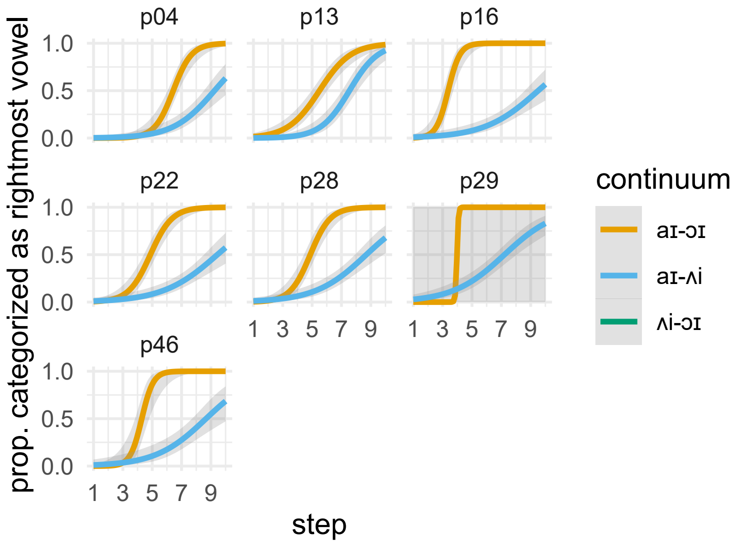


Figure 5.10: Response curves for all participants whose [a͡ɪ]-[ʌ͡i] response curve is not linear and also not sigmoidal.

distinguish between [a͡ɪ] and [ʌ͡i]. These flat responses are all generally below 0.5, so we can also note that this group, comprising 13% of all participants, were reluctant to categorize vowels as [ʌ͡i] regardless of continuum step, a result again counter to expectations based on unigram and bigram probabilities.

5.3.4 Nonlinear, non-sigmoidal responses to [a͡ɪ]-[ʌ͡i]

Finally, it should be noted that 7 individuals (participants 4, 13, 16, 22, 28, 29, and 46; 22% of participants) had blue, [a͡ɪ]-[ʌ͡i] response curves with a lower asymptote but no upper asymptote. As such, these participants, shown in Fig. 5.10, had response curves that were not fully sigmoidal but also nonlinear — that is, these curves give the impression of a sigmoid, but would require the phonetic continuum to be extended past an authentic endpoint stimulus to display the characteristic upper asymptote. As this sort of response curve is novel to me, it is difficult to interpret. On one hand, a naive suggestion would be that these participants have [ʌ͡i] as a marginal phoneme. However, Gelbart (2005) tested classification of marginal contrasts and did not find response curves of this shape. Then again, Gelbart only reported on aggregate data and I have already discussed why I am looking at individual data.

In any case, when compared to participants with fully sigmoidal response curves, there is no statistically significant difference in [a͡ɪ]-[ʌ͡i] token frequency difference measures ($t(6.3) = -1.52$, $p = .176$) and thus no difference in [ʌ͡i] token frequency or [a͡ɪ] token frequency. There was also no statistically significant difference in summed [a͡ɪ]+[ʌ͡i] type informativity ($t(13.0) = -0.10$, $p = .924$), [a͡ɪ] type informativity ($t(9.0) = -0.76$, $p = .466$) or [ʌ͡i] type informativity ($t(12.9) = 0.27$, $p = .795$). A final time, my experiment does not find evidence to support the hypothesis that lexical informativity has an effect on k.

5.4 Discussion

In this chapter, I have shown that participant behavior is orderly in that it is possible to group participants by behavior but that statistical tests generally do not find significant differences for lexical measures between groups of participants. Even when comparing groups with statistically significant differences, behavior-based groupings are not satisfactorily explained by reference to the lexical statistics. Take the single significant difference that was found: participants with complete overlap in [a͡ɪ]-[ɔ͡ɪ] and [ʌ͡i]-[ɔ͡ɪ] response curves had lower summed [a͡ɪ]+[ʌ͡i] type informativity, which should result in worse contrast between [a͡ɪ] and [ʌ͡i]. Though this finding tracks with predictions, we would also predict that low summed informativity would lead to linear response curves. Yet, of these nine participants (cf. Fig. 5.3), only five (participants 1, 17, 21, 45, 46) have linear response curves in [a͡ɪ]-[ʌ͡i] (in blue; cf. Fig. 5.8) — the other four had clearly sigmoidal response curves in [a͡ɪ]-[ʌ͡i], indicating categorical perception and thus strong contrast between [a͡ɪ] and [ʌ͡i].

Thus, the overall theme of this chapter emerges: though participants behave in patterned ways, though lexical measures are implied by the literature to support the creation of strong phonological contrasts, the patterns and the measures don't track. Lexical token frequency difference was not

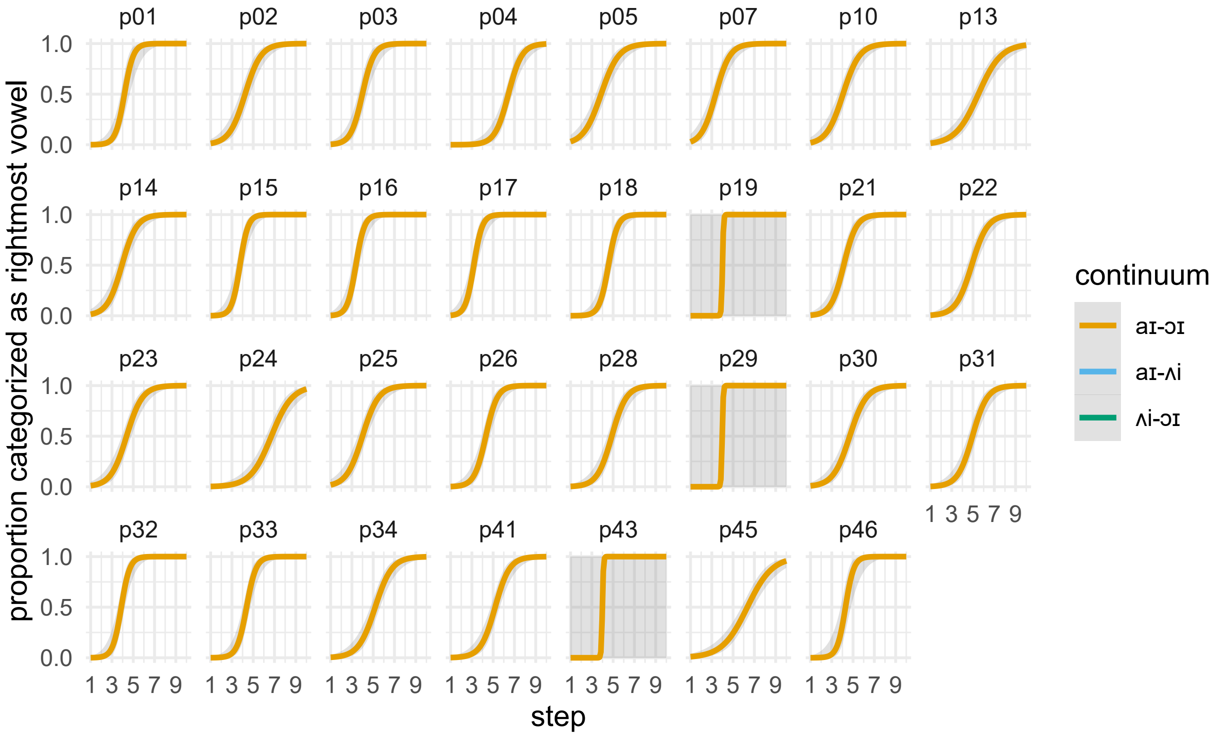


Figure 5.11: [a͡ɪ]-[ɔ͡ɪ] response curves for all participants.

found to bias participants one way or another. And summed type informativity was not found to reliably predict presence or degree of categorical perception of [a͡ɪ]-[ʌ͡i].

The disorder of these findings combines with two separate pieces of additional evidence to suggest that there is a fundamental problem with the use of these lexical measures as predictors of contrast strength.

5.4.1 Inconsistency of Extended Ganong effect

First, there is the issue of how consistently the Extended Ganong effect holds across different contrasts. Fig. 5.11 shows that nearly all participants categorize ambiguous stimuli in [a͡ɪ]-[ɔ͡ɪ] in a way that is not contrary to the Extended Ganong effect: as the unigram and bigram probabilities are higher for the [ɔ͡ɪ] nonce word than the [a͡ɪ] one, following Steffman & Sundara (2024), we expect

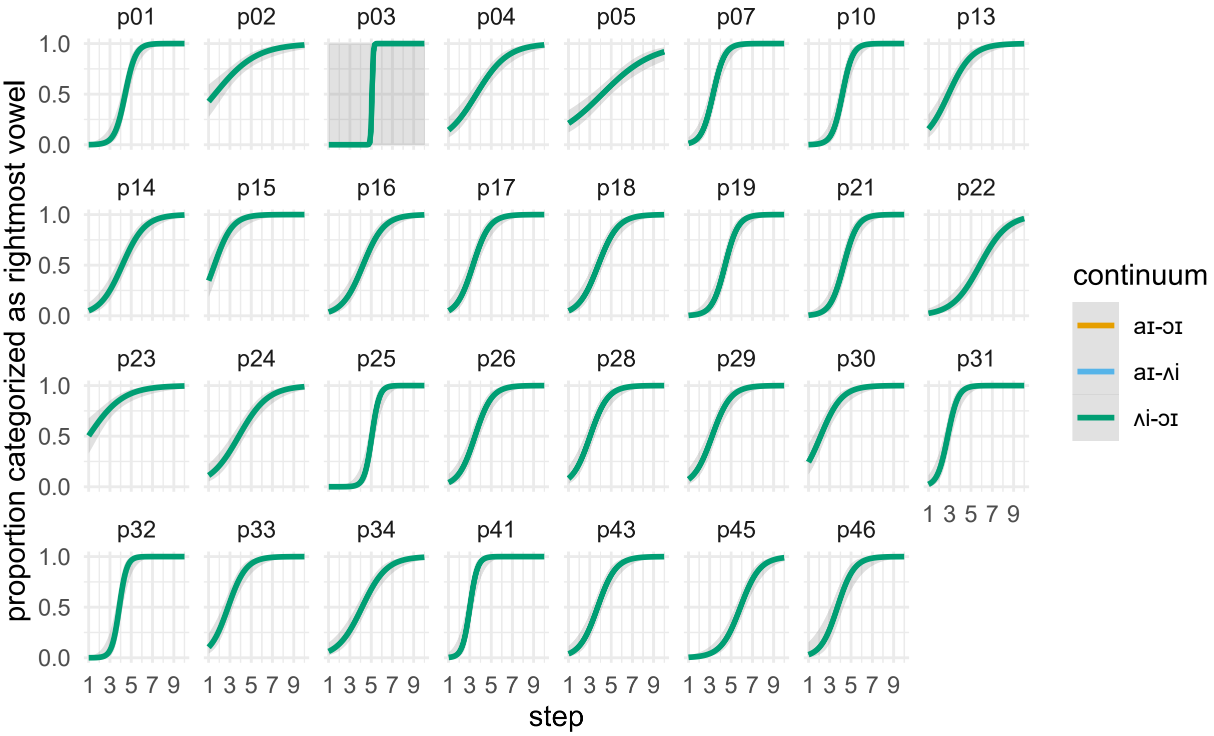


Figure 5.12: [ʌ͡i]-[ɔ͡ɪ] response curves for all participants.

to see a shift of the response curve to the left, resulting in more [ɔ͡ɪ] categorizations overall. Indeed, looking at the gold response curves, we see that only 3/31 individuals (participants 4, 24, 45) have right-shifted curves while all others have either left-shifted or centered curves.

Given the unigram and bigram probabilities, the Extended Ganong effect predicts that the [ʌ͡i]-[ɔ͡ɪ] continuum should be right-shifted, resulting in more [ʌ͡i] categorizations overall. However, Fig. 5.12 shows that all individuals but participant 45 (and perhaps 22) had either left-shifted or centered response curves. Why does Steffman & Sundara's (2024) result seem to not hold?

The most straightforward interpretation is that even the Extended Ganong effect does not apply when considering a marginal phoneme such as [ʌ͡i]. In other words, unigram and bigram probabilities do not predict a categorization shift for non-strong contrast. Indeed, as discussed in §5.2.2 and seen in Figs. 5.4 and 5.12, many participants were overall reluctant to categorize even unaltered productions of [ʌ͡i], step 1 of the green continuum in the mentioned figures, as [ʌ͡i]. In other

	sum informativity mean (sd)	frequency diff. mean (sd)
a͡ɪ-ɔ͡ɪ	7.77 (.004)	77033 (1362)
a͡ɪ-ʌ͡i	7.53 (.012)	62831 (2724)
ʌ͡i-ɔ͡ɪ	9.08 (.014)	14202 (1362)

Table 5.1: Mean and standard deviations for summed type informativity and token frequency difference for all three contrasts. Informativity is in bits and frequency is in occurrence per million words.

words, perhaps it is not licit to assume that predictions that hold true for strong phonemes will be applicable to marginal phonemes. Though speculative, this interpretation is bolstered by the fact that an independent piece of evidence supports the same conclusion.

5.4.2 Informativity is inversely related to frequency

That second piece of evidence is that the overall pattern for informativity across continua does not predict contrast strength. Recall that informativity is the average predictability of a sound and so high informativity for a vowel means that accurate categorization of that vowel is, on average, important to accurate identification of the word. This is why I posited that summed informativity should lead to faster reaction times overall, as those vowels are the most important for people to attend to while listening.

For ease of reference, I reproduce parts of Table 3.2 and Table 3.3 here as Table 5.1. Table 5.1 shows that while summed informativity is higher for the [a͡ɪ]-[ɔ͡ɪ] contrast than the [a͡ɪ]-[ʌ͡i] contrast, both contrasts have lower summed informativity than the [ʌ͡i]-[ɔ͡ɪ]. Thus, if summed informativity were the primary predictor of strength of contrast, then, *mutatis mutandis*, [ʌ͡i]-[ɔ͡ɪ] should have a stronger categorical response than [a͡ɪ]-[ɔ͡ɪ].

The reason why [ʌ͡i]-[ɔ͡ɪ] has higher summed informativity than [a͡ɪ]-[ɔ͡ɪ] comes directly from how informativity is calculated. Since infrequent sounds are less predictable relative to frequent

	informativity mean (sd)	frequency mean (sd)
a͡ɪ	3.11 (.005)	79910 (1362)
ʌ͡i	4.42 (.014)	17080 (1362)
ɔ͡ɪ	5.51 (0)	2878 (0)

Table 5.2: Mean and standard deviations for type informativity and token frequency for all three phones. Informativity is in bits and frequency is in occurrence per million words. Note that sd of token frequency is the same for [a͡ɪ] and [ʌ͡i] due to how [ʌ͡i] was implemented in the corpus (cf. §3.2) and that the sd for [ɔ͡ɪ] informativity and frequency are both 0 because participants were not assumed to vary in which words containing [ɔ͡ɪ] are in their lexicons.

sounds, the rarer a phoneme is, the more informative it will on average be. Table 5.2 shows how this played out for the participants in this experiment. We see now that the calculation of informativity of any marginal phoneme will likely result in a high value, simply because marginal phonemes tend to be less frequent than non-marginal ones. This is potentially an inevitable outcome if informativity is calculated for marginal phonemes by assuming that the marginal phoneme can be treated as a non-marginal one.

Thus, it may be the case that summed informativity and frequency difference could not have been good predictors of the eye tracking and categorization results because marginal phonemes cannot be assumed to behave like strong phonemes for the purposes of calculating these metrics. Perhaps an alternative method of calculating informativity could resolve this issue, but such a proposal is beyond the scope of this dissertation.

5.4.3 Patterns of behavior

Though the lexical measures turned out to be flawed in the end, the categorization results show that participants do fall into behavioral groups when categorizing in the [a͡ɪ]-[ʌ͡i] continuum. For convenience of discussion, the categorization results for this continuum are presented in isolation

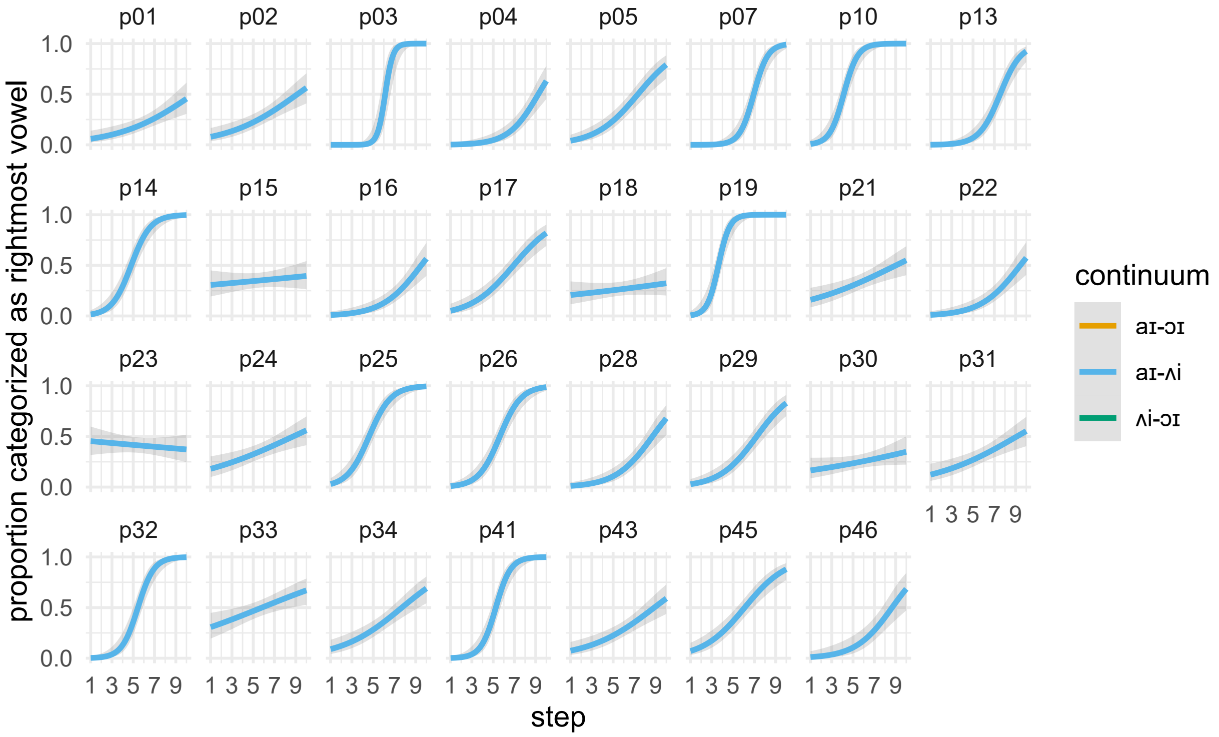


Figure 5.13: [a͡ɪ]-[ʌ͡i] response curves for all participants.

in Fig. 5.13 and Fig. 5.1, showing response curves for all continua, is repeated as Fig. 5.14.

As discussed in §5.3, 9/31 (29%) of participants had fully sigmoidal response curves, indicating the presence of categorical perception; 7/31 (22%) had nonlinear response curves that were not fully congruent with categorical perception; 11/31 (35%) had non-flat, linear response curves, indicating gradient perception; and 4/31 (13%) had flat response curves, indicating an inability to distinguish between [a͡ɪ] and [ʌ͡i].

What information predicts which group an individual will pattern with remains to be seen, though it does not appear to be either type informativity or token frequency. We can assume that participants with fully sigmoidal response curves for [a͡ɪ]-[ʌ͡i] have fully phonologized [ʌ͡i] — in fact, these 9 participants (3, 7, 10, 14, 19, 25, 26, 32, 41) had sigmoidal responses to all three continua. We can further assume that the 4 participants with flat response curves (15, 18, 23, 30) have not phonologized [ʌ͡i]. But what distinguishes those with linear, non-flat response curves

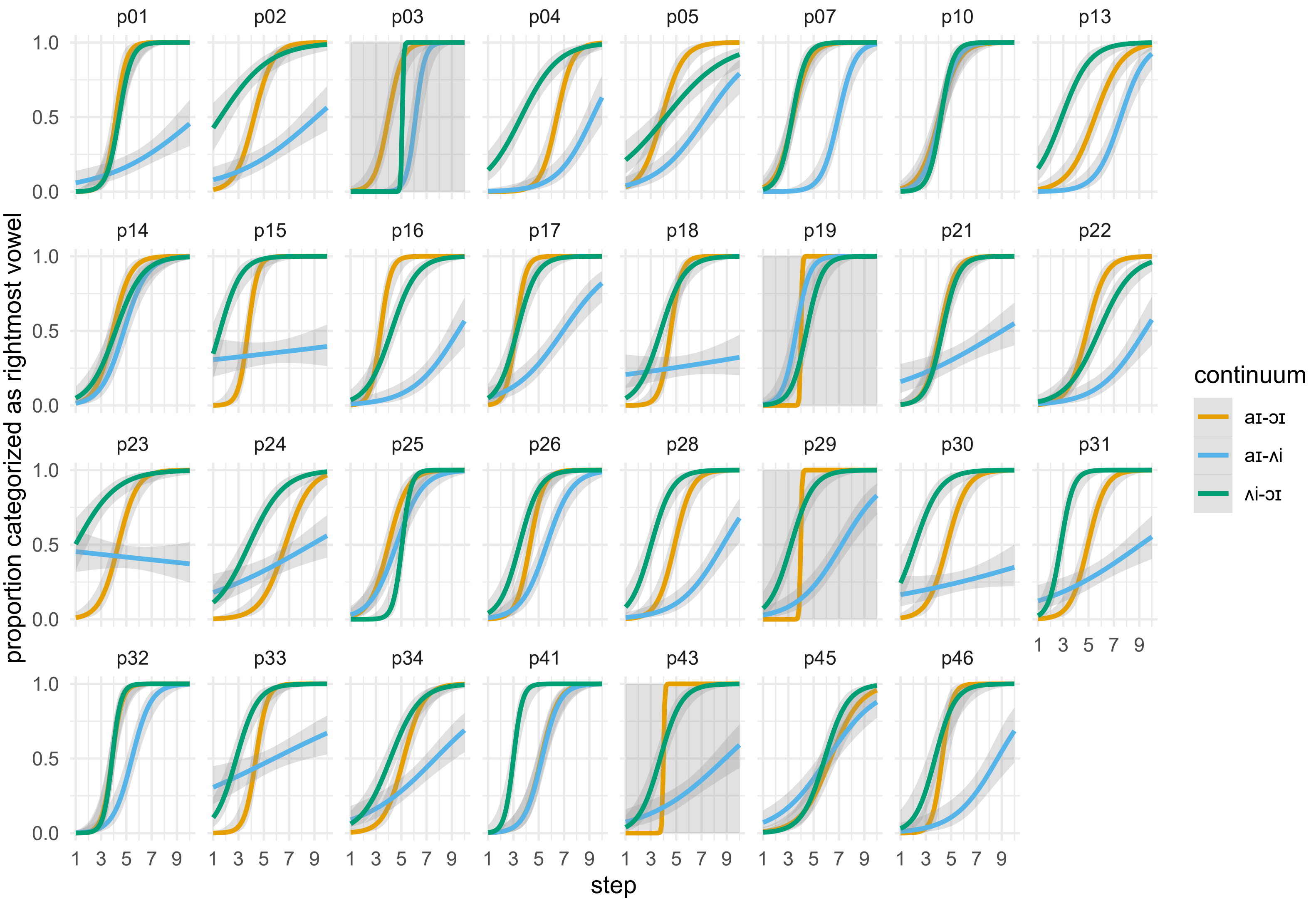


Figure 5.14: Repeat of Fig. 5.1. Response curves for all three continua for each participant. Continuum step is on the x-axis, with step 1 being the left vowel endpoint, and proportion categorization as left vowel is on the y-axis. 95% CIs are included.

from those with flat response curves? A linear response curve is generally taken to be indicative of gradient, phonetic perception, so these participants would be assumed to have not phonologized [ʌ͡i] — do these 11 participants (1, 2, 5, 17, 21, 24, 31, 33, 34, 43, 45) just have more sensitive ears than the flat responders?

More pressingly, what distinguishes the participants with nonlinear, nonsigmoidal response curves from those with sigmoidal response curves? It is tempting to label the nonlinear, nonsigmoidal responders as having a marginal [ʌ͡i] phoneme, but as mentioned in §5.3.4, there is no statistically significant difference between these groups in any of the lexical measures. These will have to be questions that future scholars take up.

5.4.4 Aggregate response curves

Finally, I will discuss the aggregate response curves of my participants and compare them to Gelbart's (2005) aggregate response curves for the Japanese /p-pp/, /d-dd/, and /b-bb/ contrasts. Note that Gelbart does not have individual response curves to report and his experimental design does not permit him to do so.[4]

In Fig. 5.15, I present Gelbart's aggregate response curves in the left panel and mine in the right panel. We can see that his response curves and mine are, in fact, decently similar. Though the aggregate response curve for [a͡ɪ]-[ʌ͡i] seems linear, careful examination shows that it is not possible to draw a straight line across the x-axis that fits entirely within the confidence interval; thus, it is slightly sigmoidal.

Arranging these curves in order of how prototypically sigmoidal they appear, the order for Gelbart's graph would be /p-pp/ > /d-dd/ > /b-bb/, and mine would be [a͡ɪ]-[ɔ͡ɪ] > [ʌ͡i]-[ɔ͡ɪ] > [a͡ɪ]-[ʌ͡i]. With the strongest contrasts at the top, a natural question is whether Gelbart's explanation for

[4] My interpretation of his procedure is that his 12 participants conducted 576 categorizations each, with each participant only encountering two out of the three continua. Since each continuum had 8 steps, each participant therefore categorized each step 36 times. Gelbart was interested in both consonant length and vowel length, so his design crossed eight steps of consonant length with six steps of following vowel length. For our purposes, we can treat the six vowel length steps as repetitions of consonant length steps.

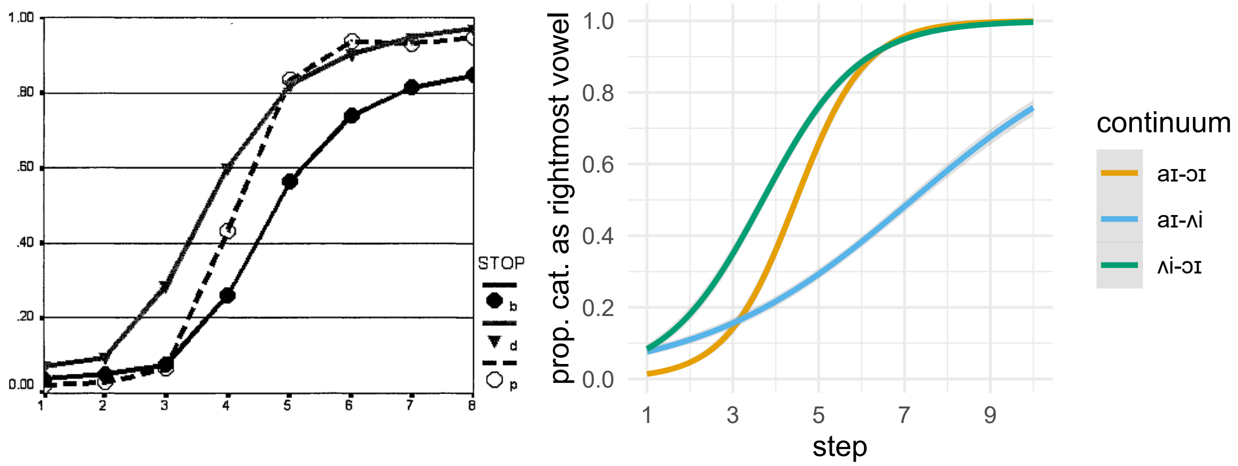


Figure 5.15: Left, Fig. 3.4 from Gelbart (2005) where y-axis is "proportion categorized as long". Right, response curves for [a͡ɪ]-[ɔ͡ɪ], [a͡ɪ]-[ʌ͡i] and, [ʌ͡i]-[ɔ͡ɪ], aggregated across participants.

this ordering also holds for the ordering of contrasts in my experiment.

Gelbart cites Amano & Kondo (2000) on the lexical frequencies of /dd/ and /bb/, but as the corpus that Gelbart cites is in Japanese (and difficult to access), I instead cite frequencies reported in Tamaoka & Makioka (2004), a follow up to Amano & Kondo (2000) using the same corpus. Thus, in Table 5.3, I present token frequencies for both members of all contrasts under discussion. I use Tamaoka & Makioka's (2004) figures for the Japanese sounds and the average token frequencies calculated in my experiment (cf. Table 5.1) for the Canadian sounds, though as the Tamaoka & Makioka (2004) figures are absolute, I also transform them to frequency per million to facilitate comparison. Further, for each contrast, I also present token frequency differences as well as the relative frequency of the less frequent member of each contrast to the more frequent. Gelbart does not say how he would go about quantifying the lexical support behind a contrast, though he wrote that "differences in frequency [between /bb/ and /dd/] may go some way to explaining the differences in [response curves]", so I present ratios merely for the sake of argument, as they pattern with the frequencies of the geminates.

Considering the numbers in Table 5.3, it is quite clear that, while the right/left ratio tracks Gelbart's results, projecting the same trend onto Canadian English would erroneously predict that

	left freq.	right freq.	freq. diff.	right/left ratio
p-pp	8037	1908	6129	.2374
d-dd	79936	79	79857	.0010
b-bb	32604	2	32602	≈ .0001
a͡ɪ-ɔ͡ɪ	77910	2878	75032	.0369
a͡ɪ-ʌ͡i	77910	17080	60830	.2192
ʌ͡i-ɔ͡ɪ	17080	2878	14202	.1685

Table 5.3: Comparison of phoneme-level and contrast-level token frequency measures for contrasts in Gelbart (2005) and this dissertation. Japanese figures are from Tamaoka & Makioka (2004) and have been normalized to be per million (instead of per 287,792,797).

[a͡ɪ]-[ʌ͡i] should have the most sigmoidal response curves and that [a͡ɪ]-[ɔ͡ɪ] should be marginal. Thus, we see that frequency ratio does not predict categorization behavior.

Unfortunately, without access to the Asahi corpus that Amano & Kondo (2000) and Tamaoka & Makioka (2004) use, it is not possible to calculate informativity for the Japanese sounds. However, as my experiment did not find informativity to be a predictor of categorization strength, and given that the rarity of /bb/ implies high informativity, I doubt that summed informativity would have predicted his results either.

Of course, Gelbart's participants surely had different lexicons with different levels of support for each of the three contrasts he examined. Moreover, all three of his continua were selected because one member of each contrast is absent from the native, Yamato stratum of Japanese — they are all marginal contrasts. In this way, his test cases were all similar to [a͡ɪ]-[ʌ͡i], so it is curious that the aggregate curves from his experiment are all clearly sigmoidal where my [a͡ɪ]-[ʌ͡i] aggregate curve is very close to linear. Of course, I have already shown in Fig. 5.13 that some participants do have a sigmoidal [a͡ɪ]-[ʌ͡i] response curve and others do not, so I can only speculate that his participants were simply more consistent than mine. Yet, if this is so, why?

Perhaps it is informative to consider why his /p-pp/ curve, a putatively marginal contrast, looks so categorical. In his conclusion, Gelbart offers three suggestions for why this may be so, but if we believe that /d-dd/, /b-bb/, and [a͡ɪ]-[ʌ͡i] are all on a cline of marginality, all three can be dismissed as unlikely. First, he suggests that the "voiced stop contrasts involve alternations, while the pp∼p contrast does not" — I take this to refer to the fact that Japanese morphophonology avoids /dd/ and /bb/ in derived environments whereas /p/ is absent from the native Yamato stratum due to historical sound changes and /pp/ is only ever the output of morphophonology (again, in Yamato). This reason could not explain [a͡ɪ]-[ʌ͡i], as [ʌ͡i] is present both in derived (e.g., *bicycle*, *high school*) and nonderived (e.g., *cider*, *ice*) environments.

A second suggestion is that the "ban on voiced geminates applies to all voiced stops, while the ban on singleton p applies only to singleton p" — this is an argument that the constraint against voiced geminates is stronger because it is broader and, while intriguing, cannot explain why [a͡ɪ]-[ʌ͡i] is so weakly sigmoidal. This argument is fundamentally about lexical strata, as this generalization only holds for the Yamato stratum, and no one to my knowledge has argued that [a͡ɪ]-[ʌ͡i] only contrast in a special stratum of Canadian English.

Finally, his last suggestion is that frequency rules the roost and the /p-pp/ contrast is stronger because /p/ is more frequent than /bb/ and /dd/ are. My experiment only considers one true marginal contrast, [a͡ɪ]-[ʌ͡i], but if we treat [ʌ͡i]-[ɔ͡ɪ] as a marginal contrast, we can see that [a͡ɪ] is much more common than [ɔ͡ɪ], suggesting that the [a͡ɪ]-[ʌ͡i] marginal contrast should evince more categorical perception than [ʌ͡i]-[ɔ͡ɪ], which it does not. However, I will readily admit that [ʌ͡i]-[ɔ͡ɪ] is not a marginal contrast in any of the ways that Hall (2013) discusses, so perhaps this equivalence is misleading. Perhaps a more convincing strike against this possibility is that the [a͡ɪ]-[ɔ͡ɪ] contrast, which we know to be strong, has a frequency profile closer to /d-dd/ than to /p-pp/.

In any case, I conclude that none of alternation presence, stratal indexation, constraint specificity, or lexical frequency are adequate to explain the differences between the /d-dd/, /b-bb/, and [a͡ɪ]-[ʌ͡i] response curves. Shifting track, we might ask why /b-bb/ is so good relative to [a͡ɪ]-[ʌ͡i]. I see two options that have not yet been ruled out, one based in lexical statistics, and one based in phonological theory. With respect to the lexicon, though I have shown that summed informativity does not predict categorization behavior, the literature still implicates predictability in contrast strength. Thus, perhaps that as-yet unknown predictability measure is sufficient. The other option is that the general support for a singleton-geminate contrast in Japanese across all lexical strata is sufficient to bring /b-bb/ and /d-dd/ to the near edge of marginality. Tamaoka & Makioka (2004) report non-zero token frequency of geminates for all consonants except for /m,n,w,j/ — that is, of the 14 indisputably phonemic singleton consonants of Japanese, 70% show some support, marginal or otherwise, for a length contrast. If vowel length is representationally the same as consonant length, then 15/19 or 79% support a length contrast. In comparison, of the 13 indisputably phonemic vowels of Canadian English, only 2 (15%) have any support for a raising contrast and the outlook only worsens if we include consonants. This argument is reminiscent of both Maddieson's (1985) observation that languages tend to borrow sounds that "fill in gaps" and Clements's (2003) principle of feature economy — perhaps future work examining the categorization of marginal contrasts in Cairene Arabic, which also seem to be gap-filling, will be able to provide evidence one way or another.

Finally, we should consider other reasons beyond the statistical for why [a͡ɪ]-[ʌ͡i] would be less categorically perceived than the Japanese marginal contrasts Gelbart considers. For one, since Fry et al. (1962), it has been generally held that vowel perception is less categorical than consonant perception, so perhaps that is to blame — though as my all participants in my experiment had sigmoidal [a͡ɪ]-[ɔ͡ɪ] response curves, it would be irregular to cite Fry et al. (1962) to explain only [a͡ɪ]-[ʌ͡i]. Bruce Hayes (p.c.) suggests an interpretation where cosmopolitan speakers of Canadian English are inundated with American English diphthongs over the course of their lives and, even if they had successfully acquired a truly phonemic /ʌ͡i/ in early childhood, correspondingly alter their perception grammars in the vein of Boersma & Hamann (2008) to deemphasize the perceptual distinctiveness of [a͡ɪ]-[ʌ͡i]. In comparison, the Japanese speakers in Gelbart's study would have never encountered any other Japanese speakers that caused them to perceptually deeemphasize a length contrast — though the production of voicing for voiced geminates in Japanese is unreliable, duration is not (Kawahara, 2005; Hussain & Shinohara, 2019), not even for liquid geminates Morimoto (2020).

Since Boersma & Hamann (2008) predicts that a speaker's production grammar would not merge /a͡ɪ/ and /ʌ͡i/ even if their perception grammar merged the two, Hayes's interpretation has the added benefit of potentially rationalizing why the lexical measures, which I estimated by asking speakers about their *production*, were poor predictors of categorization behavior. Without information on participant life histories and media consumption, it is not possible to confirm or deny the applicability of this account, though it makes the prediction that marginal phonemes, whether gap-filling or otherwise, should elicit more categorical perception if prominent varieties in contact have that marginal phoneme as a strong phoneme.

5.5 Chapter summary

In this chapter, I have shown that the categorization results pattern in ways that are not predicted by summed type informativity, token frequency difference, or the constituent phoneme-level measures that these contrast-level measures are composed of.

Having established the behavioral groups that I observe in the data, I discussed why I believe the lexical predictors I used are flawed, namely that the act of calculating them assumes phonemicity in a potentially problematic way. I pointed to both the falsified prediction that Steffman & Sundara's (2024) Extended Ganong effect would right-shift the [ʌ͡i]-[ɔ͡ɪ] response curves and the falsified prediction made by summed informativity that ambiguous stimuli in [ʌ͡i]-[ɔ͡ɪ] should elicit a more categorical response than those in [a͡ɪ]-[ɔ͡ɪ] as evidence that extending predictions for strong contrasts to marginal contrasts may be inappropriate.

Further research on what information within the mind of the individual can successfully predict contrast strength will have to be done. The results from this chapter raise two related questions for future scholars to consider:

- What predicts the strength of an individual's reluctance to categorize naturalistic stimuli of marginal phonemes as said phoneme?
- What predicts whether an individual will have a phonologized "marginal" contrast, a true marginal contrast, or no phonological contrast?

Of course, these research questions only hold up if indeed the variation we see within each group (sigmoidal, nonlinear, etc.) is meaningful. Thus, I suggest it would be ideal for future researchers to first tackle the more specific questions below:

- For individuals with a phonologized marginal contrast, what predicts the presence of the (Extended) Ganong effect?
- For individuals with a nonlinear, non-sigmoidal response curve (when categorizing stimuli along a phonetic continuum representing a marginal contrast), what predicts the degree of reluctance to categorize naturalistic stimuli of marginal phonemes as said phoneme?
- For individuals with a linear response curve, what predicts the slope of the response curve?

CHAPTER 6

Conclusion

I began this dissertation with the hope of producing evidence that phoneme strength is both predicted by individual lexicons and in turn predicts behavior for tasks that have been classically explained by reference to phonemic category status. This hope, primarily motivated by the lurking unease engendered in me by the ill-definedness of a “marginal contrast”, resulted in my identification of frequency and predictability as primary concerns when scholars have called a contrast marginal, identifying token frequency and type informativity as reasonable operationalizations of frequency and predictability, and positing that token frequency and type informativity would have an impact on behavior during a category-based task such as a categorization task.

Thus, I designed and conducted an appropriate study on the categorization of stimuli across the [a͡ɪ]-[ʌ͡i], [a͡ɪ]-[ɔ͡ɪ], and [ʌ͡i]-[ɔ͡ɪ] continua by Canadian English speakers. Analyzing the resulting data, what I expected to see was generally categorical perception of the marginal [ʌ͡i] phoneme, moderated by lexical frequency and informativity. Instead, we saw that while participant behavior can be grouped, group membership was not predicted by my selected metrics. While model comparison shows that token frequency difference and summed type informativity were statistically significant predictors of looking bias, the uninterpretability of these partial effects, the lack of correlation with behavioral groups, and converging evidence from Steffman & Sundara’s (2024) Extended Ganong effect and general expectations for how stimuli from the [a͡ɪ]-[ɔ͡ɪ] would be received by participants all suggest that these predictors were not ideal for my intended purpose.

At this point in my dissertation, I want to return to a brief discussion of the nature of science. Though my original hope for this dissertation has not borne the fruit I sought, neither has effort been wasted. I do not believe the chain of logic I presented in chapter 2 is incorrect: within the literature, predictability and frequency *are* implicated in the construction of phonological categories by both theoretical prediction and experimental outcome. Thus are we positioned similarly to how we were at the start of this project, caught in a different contradiction, but caught nonetheless. Yet the scientific endeavor is not driven by positive evidence alone, so progress has been made.

My positive contributions to the field are the first ever categorization study of a marginal phoneme with an unrelated strong phoneme and empirical finding that participant behavior in categorization of a marginal contrast is not random but ordered, though it is unclear what principle(s) are responsible. My negative contribution is that summed type informativity and token frequency difference did not predict participant behavior.

I am compelled to first discuss my negative contribution: why did informativity and token frequency not predict participant behavior? I have discussed in chapter 5 the potential conceptual problem with calculating these lexical statistics and strongly suspect that this is to blame. If correct, then further exploratory work will be necessary to identify the correct lexical statistics that predict behavior: I suggest that future scholars consider their operationalization options expansively and reanalyze my data with different lexical statistics.

The issue may also lie in choice of [ʌ͡i] as primary object of inquiry. Early in chapter 2, I wrote, “tapping [...] causes the voicing contrast to shift from being redundantly evinced on both vowel quality and presence of voicing to vowel quality alone”. On this description alone, there is something funny about the finding that the informativity of [ʌ͡i] is greater than that of [ɔ͡ɪ]: shouldn’t the informativity of [ʌ͡i] be, in some sense, spread out across both the vowel and the following consonant?[1] It may be that a derived marginal contrast like [a͡ɪ]-[ʌ͡i] should be instead considered as a [a͡ɪɾ]-[ʌ͡iɾ] contrast and informativity should be calculated as if these sequences are one unit, though, because exceptional [ʌ͡i]/[a͡ɪ] is not limited to the pre-tap environment, I do not know what the implications are for proper calculation of informativity of [ʌ͡i].

[1] Indeed, when ordering all sounds present in `CMU-SUB_raised` (cf. §3.2) by informativity, [ʌ͡i] is the sixth most informative at 4.415 bits, coming in only after [a͡ʊ] (4.928), [ð] (4.714), [ɔ͡ɪ] (4.662), [h] (4.458), and [θ] (4.431).

That said, my findings have lead me to believe that current information theoretic approaches to phonetics and phonology are likely overly idealized/overly discretized. Given that Canadian Raising *is* productive, the most accurate measure of the informativity of [ʌ͡i] would surely take into account following context as well as preceding. But when PCT calculates the informativity of a sound, it only considers the preceding context (cf. Cohen Priva, 2015). Thus, I suspect a foundational error with my experiment was the assumption that disregarding following context in the calculation of informativity for [ʌ͡i] was an acceptable idealization — most shifted contrast marginal phonemes would have likely run into this same issue.[2]

As the link between the lexicon and categorization behavior ought to hold for all contrasts, it may be thus prudent to conduct a follow-up study on the impact of frequency and predictability on a set of strong phonemes before looking at further marginal phonemes. For example, my experiment could be repeated with Californian English on the [ʊ]-[ɪ], [ɪ]-[ə], and [ʊ]-[ə] continua,[3] though a different method of estimating individual lexicons would have to be employed.

Looking at my positive contributions to the field, my [ʌ͡i]-[ɔ͡ɪ] categorization task is, I believe, the first such task that has been done, and my comparison of [ʌ͡i]-[ɔ͡ɪ] and [a͡ɪ]-[ɔ͡ɪ] response curves represents a novel method of comparing the results of categorization experiments without a norming study — though the comparison is less direct than if I had been able to determine JND for each participant for each continuum, the time saved by this approach is substantial.

Further, the response curves of individual participants for [a͡ɪ]-[ʌ͡i] are also novel and the patterns of variation within behavioral groups represent a series of mysteries to be solved. Questions relating to these curves were given a mere three pages ago, so I will not repeat them here. However, as I said in chapter 2, I do believe that the field has been historically overeager to describe languages

[2] This is not to cast blame on scholars in the field, as the foundational work on information theory (e.g., Shannon, 1948, Hamming, 1950) is on discrete, symbolic systems and, given the importance of digital (binary) systems to modern society, of course that type of research is in high demand. Still, I believe phoneticians and phonologists, as scholars that must contend with the analog communication channel of the acoustic signal, are extremely well positioned to take on the work of generalizing information theory to less discrete systems — for example, see Iskarous et al. (2013).

[3] I suggest Californian English because Californian [ɪ, ʊ, ə] are quite acoustically similar, likely more similar than they are in other accessible varieties of English.

as homogeneous systems, and I suspect that increased attention to individuals, not aggregates, may be a productive way of moving forward on thorny issues. Though making sense of individual behavior can be time-consuming — and the behavior itself, noisy — as scientists, we cannot be afraid of work that is plainly yet to be done.

In this last, final chapter of my dissertation and the preceding chapter, I have given some suggestions for experiments that I believe can build on the results and null results of my dissertation. I hope that these new directions will be productive ones and, girded with that new knowledge, future scholars will be able to project their understanding back onto the questions that were at the heart of this project (and, if I may, close to the heart of the field): what makes a contrast marginal? what makes a contrast contrastive?

APPENDIX A

Experiment instructions

Training portion

In this task, you will learn to associate three pictures of objects with three words: kliter, klider, and kloiter. It is your job to learn which picture goes with which word. At the beginning, you may have to guess, but you'll receive feedback on your responses and you'll learn over time. This portion of the experiment will take about 5 minutes.

Click on the circle in the center of the screen to hear the word, then click on the picture you think matches the word you heard. You will see a green box around the correct answer.

Eyetracking portion

Calibration instructions: We will now calibrate the eyetracker. Please put the sticker on your forehead, between and slightly above your eyes. Sit in a comfortable position so that you can reach the mouse and see the screen. Now we have to teach the eyetracker what it looks like when your eyes look at different parts of the screen. Please follow the dot with your eyes as it moves around the screen. We will do this a few times as we calibrate the tracker.

Experiment instructions: In this portion of the experiment, you will be tested on the words and images you learned in the training. You will see two images on the screen with a red circle in the center. Look at the red circle until it turns blue, then click on the blue circle to start the trial. You will hear one of the words you learned earlier. Click on the matching picture. The experiment will then automatically move on to the next trial. Sometimes you may not be sure which word you heard—that's okay, just make your best guess and move on. The computer will record where your eyes are looking on the screen as you do the task. Use natural eye movements and don't be afraid to blink. Every 30 or so trials, you will see a break screen. When you see that screen, feel free to stretch, take a drink of water, etc. When you are ready to continue, look at the dot in the center of the screen and press the spacebar. This helps us confirm that the eyetracker still has a good calibration. I will be monitoring the experiment from the room next door. You can see me through the window. If you have any trouble or need anything, just wave and I'll come in to help. This portion of the experiment will take about 45 minutes.

Survey portion

This survey asks about your language background and your pronunciation of some English words. Please answer honestly. This portion of the experiment will take about 20 minutes. You may choose to stay in the lab to complete the survey, or we can email you a link to do it at home. If you choose to do it at home, please be sure that you are listening on headphones in a quiet room.

Survey instructions

We want to learn how you pronounce a few words. For each word, decide whether *you* pronounce the dark red vowel more like the vowel in **writer** or more like the vowel in **rider**. A recording of both pronunciation options will be on each page and you may play them as many times as you like.

Do not worry about "correct" pronunciation or phonological rules; simply consider how you say the word. You may want to say the words out loud, but do not ask others for their judgments.

Please use wired headphones/speakers for this portion of the experiment as wireless devices may encounter playback issues.

APPENDIX B

Exhaustive exception list

This is the `exhaustive_exception_list` described in §3.1.4 that I used to create the survey we gave to participants after completion of the eye tracking portion of the experiment. In this list, the headword is given first, followed by all words we considered derived from that headword (but not repeating the headword itself).

In the survey, we asked participants about the pronunciation of headwords and propogated the vowel they reported in the headword to derived words. For example, the corpus for a participant who indicated that they pronounce *dine* as [dʌ͡in] would have [dʌ͡in, dʌ͡ind, dʌ͡inz, dʌ͡inɪŋ].

Individual survey responses are all available at the supplementary materials at osf.io/4xveb under `survey files>responses`. In these responses, the format is `"headword":"#X"`, where X=1 indicates an [a͡ɪ] response and X=2 indicates an [ʌ͡i] response. The resultant corpora used to calculate lexical statistics are under `survey files>corpus files>individual corpora`, and `survey files>summary.csv` contains the calculated lexical statistics.

1. bicarbonate
2. biceps
3. bipolar
4. bisexual: bisexuals
5. cider
6. citation: citations
7. cite: cited, cites, citing
8. cyclops
9. desire: desired, desires, desiring
10. dine: dined, dines, dining
11. diner: diners

12. dissect: dissected, dissecting, dissection, dissections, dissects

13. entire: entirely

14. fire: bonfire, bonfires, campfire, campfires, fiery, fireballs, fired, firefight, firefighter, firefighters, firefighting, firefights, firehouse, firehouses, fireman, firemen, fireplace, fireplaces, fireproof, fireproofing, fires, firewood, firing, firings, wildfire

15. friday: fridays

16. gigantic

17. hire: hired, hires, hiring

18. hype: hyped, hyper, hypes, hyping

19. hypothesis: hypotheses

20. icon: icons

21. idle: idled, idler, idles, idling, idly

22. inquire: inquired, inquires, inquiring

23. inspire: inspired, inspires, inspiring

24. iris: irises

25. irish

26. iron: irons

27. irony

28. life

29. like

30. nice: nicely, niceness, nicer, nicest

31. nine: nines, nineteen, nineteenth, nineties, ninetieth, ninety, ninth

32. nitrate: nitrates

33. pint: pints

34. pirate: pirated, pirates, pirating

35. psychology

36. psychotic

37. python

38. siberia: siberian

39. sire: sires

40. spider: spiders

41. spiral: spiraled, spiraling, spiralling, spirals

42. tiger: tigers

43. tire: tired, tires, tiring

44. titanic

45. trifecta

46. tripod: tripods

47. tycoon: tycoons

48. typhoon: typhoons

49. vicarious: vicariously

50. vitality

51. wire: wired, wires, wiring, wiry

APPENDIX C

Headword responses by participant

This appendix reports participant responses to the pronunciation survey. Participants who were excluded from the study are not included here — all participants who were excluded either did not complete the study, were not a native speaker of Canadian English, or incorrectly answered at least one catch question.

The columns are participants, rows are headwords — cf. Appendix B and §3.1.4, and cells contain the diphthong which the participant reported having in the headword. For the sake of visual clarity, "." stands for [a͡ɪ] and "ʌ", [ʌ͡i].

This information is also accessible as a .csv file in the supplementary materials at osf.io/4xveb as `exceptions_summary.csv`. In that file, 0 indicates an [a͡ɪ] and 1 indicates an [ʌ͡i].

	p01	p02	p03	p04	p05	p07	p10	p13	p14	p15	p16	p17	p18	p19	p21	p22
bicarbonate	ʌ	.	.	.	.	.	.	.	ʌ	.	ʌ	.	.	.	ʌ	.
biceps	ʌ	.	.	.	ʌ	.	.	.	.	.	.	.	.	.	.	.
bipolar	ʌ	.	ʌ	.	ʌ	.	.	.	.	.	ʌ	ʌ	.	.	ʌ	.
bisexual	.	.	.	ʌ	ʌ	.	.	.	.	.	.	ʌ	.	.	ʌ	.
cider	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
citation	ʌ	ʌ	ʌ	ʌ	.	ʌ	ʌ	.	ʌ	ʌ	ʌ	ʌ	ʌ	ʌ	ʌ	.
cite	ʌ	ʌ	ʌ	ʌ	.	ʌ	ʌ	.	ʌ	ʌ	ʌ	ʌ	ʌ	ʌ	ʌ	.
cyclops	.	.	ʌ	ʌ	.	ʌ	.	.	.	.	.	.	.	ʌ	.	ʌ
desire	ʌ	.	.	.	.	.	.	ʌ	.	.	.	.	.	.	.	.
dine	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
diner	.	.	.	.	.	.	.	ʌ	.	.	.	.	.	.	.	.
dissect	ʌ	.	ʌ	.	.	.	ʌ	.	.	ʌ	ʌ	.	ʌ	ʌ	ʌ	.
entire	.	.	.	.	ʌ	.	.	.	.	ʌ	.	ʌ	ʌ	.	ʌ	ʌ
fire	.	.	.	ʌ	.	.	ʌ	.	.	.	.	.	.	.	.	.
friday	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
gigantic	ʌ	ʌ	ʌ	.	ʌ	.	.	.	.	.	ʌ	ʌ	ʌ	.	ʌ	.
hire	.	.	.	.	ʌ	ʌ	.	.	.	.	.	ʌ	.	.	.	.
hype	ʌ	ʌ	ʌ	ʌ	ʌ	ʌ	.	.	.	ʌ	ʌ	ʌ	ʌ	ʌ	.	ʌ
hypothesis	ʌ	.	.	ʌ	.	.	.	.	.	.	ʌ	ʌ	.	.	ʌ	ʌ

Continued from previous page

	p01	p02	p03	p04	p05	p07	p10	p13	p14	p15	p16	p17	p18	p19	p21	p22
icon	.	∧	.	.	.	.	.	.	∧	.	∧	∧	.	.	.	.
idle	.	.	.	.	.	.	.	.	.	.	∧	.	.	.	∧	∧
inquire	∧	.	.	.	.	.	.	∧	.	.	.	.	∧	.	.	.
inspire	.	.	.	∧	∧	.	.	.	.	.	∧	.	.	.	.	.
iris	.	.	∧	∧	.	.	.	∧	.	.	.	.	∧	.	.	.
irish	.	.	.	.	∧	.	.	.	.	.	.	.	.	.	∧	∧
iron	.	.	.	.	∧	.	.	.	.	.	.	.	.	.	.	.
irony	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
life	∧	.	∧	∧	∧	∧	∧	.	∧	.	∧	∧	∧	∧	∧	.
like	∧	∧	∧	∧	∧	∧	∧	.	∧	∧	∧	∧	∧	∧	∧	.
nice	∧	∧	∧	∧	.	∧	∧	.	∧	∧	∧	∧	∧	.	∧	∧
nine	∧	.	.	.	∧	.	.	.	.	∧	.	∧	.	.	∧	.
nitrate	∧	.	∧	∧	.	∧	∧	.	.	∧	∧	∧	∧	∧	∧	.
pint	.	.	.	.	.	.	.	.	.	.	∧	∧	∧	.	∧	∧
pirate	.	∧	∧	∧	.	.	∧	.	.	.	.	.	∧	.	.	.
psychology	∧	∧	∧	∧	∧	∧	∧	∧	∧	∧	∧	∧	∧	∧	∧	∧
psychotic	∧	.	∧	∧	∧	.	.	.	∧	∧	∧	∧	∧	∧	∧	.
python	∧	.	∧	.	∧	.	∧	.	.	∧	.	∧	∧	∧	.	.

Continued from previous page

	p01	p02	p03	p04	p05	p07	p10	p13	p14	p15	p16	p17	p18	p19	p21	p22
siberia	∧	.	.	.	.	.	.	.	.	∧	∧	∧	∧	.	.	.
sire	.	.	.	.	.	.	.	∧	.	.	.	.	.	.	.	.
spider	.	.	∧	.	.	∧	∧	.	.	.	∧	∧	∧	.	.	.
spiral	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
tiger	∧	.	.	.	.	∧	.	∧	.	.	.	∧	∧	.	∧	∧
tire	∧	.	.	.	.	.	.	.	.	.	.	∧	∧	.	∧	.
titanic	∧	.	∧	∧	∧	.	.	∧	.	.	∧	∧	.	.	.	.
trifecta	.	.	∧	∧	.	.	.	.	.	.	∧	.	.	.	∧	.
tripod	∧	∧	.	.	.	.	.	∧	.	.	∧	.	.	.	∧	.
tycoon	∧	.	.	∧	.	.	∧	∧	.	∧	∧	.	∧	.	.	∧
typhoon	.	.	.	∧	.	∧	.	.	.	.	∧	.	∧	.	.	∧
vicarious	∧	.	∧	.	.	.	.	.	.	.	∧	∧	.	.	.	.
vitality	.	∧	∧	.	.	.	.	.	.	.	∧	∧	.	∧	∧	.
wire	∧	.	.	.	.	.	.	.	.	.	.	.	∧	.	.	.

	p23	p24	p25	p26	p28	p29	p30	p31	p32	p33	p34	p41	p43	p45	p46
bicarbonate	.	Λ	Λ	.	.	.	Λ	.	.	.	.	.	.	.	.
biceps	.	.	Λ	.	.	.	Λ	.	.	Λ	.	.	Λ	.	.
bipolar	.	.	Λ	.	Λ	Λ	.	.	.	Λ	.	.	.	.	Λ
bisexual	.	Λ	.	Λ	Λ	.	.	.	.	Λ	.	.	.	.	.
cider	.	.	.	.	.	.	.	.	.	.	.	.	Λ	.	.
citation	.	Λ	.	Λ	.	.	Λ	Λ	Λ	Λ	.	.	Λ	.	Λ
cite	Λ	Λ	Λ	Λ	Λ	.	.	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ
cyclops	Λ	.	Λ	Λ	.	Λ	.	.	.	Λ	.	Λ	Λ	Λ	.
desire	.	.	.	.	Λ	.	.	.	.	.	.	.	.	.	.
dine	.	.	Λ	.	.	.	.	.	.	.	.	.	.	.	.
diner	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
dissect	.	Λ	.	Λ	.	Λ	.	Λ	Λ	Λ	.	Λ	.	.	Λ
entire	.	Λ	.	Λ	Λ	.	.	Λ	Λ	.	.	.	.	Λ	.
fire	.	Λ	Λ	.	.	.	.	Λ	.	.	.	Λ	.	.	.
friday	.	.	.	Λ	.	.	.	.	Λ	.	.	.	.	.	Λ
gigantic	Λ	Λ	Λ	.	.	.	.	Λ	Λ	Λ	Λ	.	.	.	.
hire	.	.	.	.	.	.	Λ	.	Λ	.	.	Λ	.	.	.
hype	Λ	Λ	Λ	Λ	.	.	Λ	Λ	.	Λ	Λ	Λ	Λ	.	Λ
hypothesis	.	.	.	.	.	.	Λ	.	.	.	.	.	.	.	.

Continued from previous page

	p23	p24	p25	p26	p28	p29	p30	p31	p32	p33	p34	p41	p43	p45	p46
icon	.	.	Λ	.	.	.	Λ	.	.	Λ	.	Λ	.	.	.
idle	Λ	Λ	.	.	Λ	.	Λ	Λ	Λ	.	.	Λ	.	.	Λ
inquire	.	Λ	.	Λ	.	.	.	.	.	Λ	.	Λ	.	.	.
inspire	.	Λ	Λ	.	.	.	.	.	.	.	.	.	.	.	Λ
iris	.	Λ	Λ	Λ	.	.	.	.	.	Λ	.	Λ	.	.	Λ
irish	.	.	.	.	Λ	.	Λ	.	.	.	Λ	.	.	.	Λ
iron	.	Λ	Λ	Λ	Λ	.	.	.	.	Λ	Λ	Λ	.	Λ	Λ
irony	.	.	Λ	Λ	Λ	.	Λ	.	Λ	Λ	.	.	Λ	.	Λ
life	Λ	Λ	.	Λ	Λ	Λ	.	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ
like	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ
nice	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ
nine	.	.	.	.	.	.	.	.	Λ	Λ	.	.	.	.	.
nitrate	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	.	Λ	Λ	.	Λ	Λ
pint	.	.	Λ	.	.	Λ	Λ	Λ	.	.	Λ	.	.	.	Λ
pirate	Λ	.	.	.	.	Λ	.	.	.	Λ	.	.	Λ	.	.
psychology	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	.	.	Λ	Λ	Λ	Λ	Λ
psychotic	Λ	Λ	Λ	Λ	Λ	Λ	Λ	Λ	.	Λ	Λ	.	Λ	Λ	Λ
python	Λ	.	Λ	Λ	.	.	Λ	.	.	.	.	Λ	.	Λ	.

Continued from previous page

	p23	p24	p25	p26	p28	p29	p30	p31	p32	p33	p34	p41	p43	p45	p46
siberia	.	.	.	.	.	Λ	.	.	.	Λ	.	.	Λ	.	Λ
sire	.	.	Λ	.	.	.	Λ	.	.	.	.	Λ	.	.	.
spider	.	.	Λ	Λ	Λ	.	.	.	Λ	.	.	.	Λ	Λ	Λ
spiral	.	.	.	.	.	.	.	.	.	.	.	Λ	.	.	.
tiger	.	Λ	Λ	Λ	.	.	Λ	.	.	.	.	.	.	.	Λ
tire	.	.	Λ	.	.	.	Λ	.	.	.	.	Λ	.	Λ	.
titanic	.	Λ	.	Λ	.	.	Λ	Λ	Λ	.	.	Λ	.	.	Λ
trifecta	.	Λ	Λ	.	Λ	.	Λ	Λ	.	.	.	.	.	.	.
tripod	.	.	Λ	.	.	.	Λ	.	.	Λ	.	.	.	.	.
tycoon	Λ	Λ	Λ	Λ	.	Λ	.	Λ	.	.	.	.	.	.	.
typhoon	Λ	Λ	Λ	.	.	Λ	.	Λ	Λ	Λ	.	Λ	.	.	.
vicarious	Λ	.	.	.	.	.	.	Λ	.	Λ	Λ	.	.	.	.
vitality	Λ	Λ	Λ	.	.	.	Λ	Λ	Λ	Λ	.	Λ	Λ	Λ	Λ
wire	.	.	Λ	.	.	.	.	.	Λ	Λ	.	Λ	.	.	.

APPENDIX D

CMU-SUB_raised features

This feature system is based on the Hayes feature system which comes with Phonological Corpus Tools. The only featural difference between `AY` [a͡ɪ] and `VY` [ʌ͡i] is that `VY` is [−low].

To import this feature system:

1. Strip out the spaces in the following column names (symbol, anterior, approximant, back, consonantal, constricted glottis, continuant, coronal, delayed_release, diphthong, distributed, dorsal, front, front-diphthong, high, labial, labiodental, lateral, long, low, nasal, round, segment, sonorant, spread glottis, stress, strident, syllabic, tap, tense, trill, voice)
2. Paste the column names into a text file and enter a new line
3. Paste the block below:

```
AE,0,+,-,-,-,+,-,0,-,0,+,+,0,-,-,-,-,-,+,-,-,+,+,-,-,0,+,-,0,-,+

B,0,-,0,+,-,-,-,-,0,0,-,0,0,0,+,-,-,-,0,-,-,+,-,-,-,0,-,-,0,-,+

P,0,-,0,+,-,-,-,-,0,0,-,0,0,0,+,-,-,-,0,-,-,+,-,-,-,0,-,-,0,-,-

L,+,+,0,+,-,+,+,0,0,-,-,0,0,0,-,-,+,-,0,-,-,+,+,-,-,-,-,-,0,-,+

ZH,-,-,0,+,-,+,+,+,0,+,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,+,-,-,0,-,+

AO,0,+,+,-,-,+,-,0,-,0,+,-,0,-,+,-,-,-,-,-,+,+,+,-,-,0,+,-,-,-,+

HH,0,-,0,-,-,+,-,+,0,0,-,0,0,0,-,-,-,-,0,-,-,+,-,+,-,0,-,-,0,-,-

JH,-,-,0,+,-,-,+,+,0,+,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,+,-,-,0,-,+

Y,0,+,-,-,-,+,-,0,0,0,+,+,0,+,-,-,-,-,-,-,-,+,+,-,-,0,-,-,+,-,+
UH,0,+,+,-,-,+,-,0,-,0,+,-,0,+,+,-,-,-,-,-,+,+,+,-,-,0,+,-,-,-,+
TH,+,-,0,+,-,+,+,+,0,+,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,-,-,-,0,-,-
SH,-,-,0,+,-,+,+,+,0,+,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,+,-,-,0,-,-
K,0,-,0,+,-,-,-,-,0,0,+,0,0,+,-,-,-,-,-,-,-,+,-,-,-,0,-,-,0,-,-
R,-,+,0,-,-,+,+,0,0,+,-,0,0,0,-,-,-,-,0,-,-,+,+,-,-,-,-,-,0,-,+
M,0,-,0,+,-,-,-,0,0,0,-,0,0,0,+,-,-,-,0,+,-,+,+,-,-,0,-,-,0,-,+
EH,0,+,-,-,-,+,-,0,-,0,+,+,0,-,-,-,-,-,-,-,-,+,+,-,-,0,+,-,-,-,+
W,0,+,+,-,-,+,-,0,0,0,+,-,0,+,+,-,-,-,-,-,+,+,+,-,-,0,-,-,+,-,+
AH N,+,-,0,+,-,-,+,0,0,-,-,0,0,0,-,-,-,-,0,+,-,+,+,-,-,-,+,-,0,-,+
OW,0,+,+,-,-,+,-,0,+,0,+,-,-,-,-,-,-,-,-,-,+,+,+,-,+,0,+,-,+,-,+
NG,0,-,-,+,-,-,-,0,0,0,+,+,0,+,-,-,-,-,-,+,-,+,+,-,-,0,-,-,0,-,+
S,+,-,0,+,-,+,+,+,0,-,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,+,-,-,0,-,-
OY,0,+,+,-,-,+,-,0,+,0,+,-,+,-,-,-,-,-,-,-,+,+,+,-,+,0,+,-,-,-,+
AH L,+,+,0,+,-,+,+,0,0,-,-,0,0,0,-,-,+,-,0,-,-,+,+,-,-,-,+,-,0,-,+
D,+,-,0,+,-,-,+,-,0,-,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,-,-,-,0,-,+
EY,0,+,-,-,-,+,-,0,+,0,+,+,+,-,-,-,-,-,-,-,-,+,+,-,+,0,+,-,+,-,+
DH,+,-,0,+,-,+,+,+,0,+,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,-,-,-,0,-,+
Z,+,-,0,+,-,+,+,+,0,-,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,+,-,-,0,-,+
IY,0,+,-,-,-,+,-,0,-,0,+,+,0,+,-,-,-,-,-,-,-,+,+,-,-,0,+,-,+,-,+
IH,0,+,-,-,-,+,-,0,-,0,+,+,0,+,-,-,-,-,-,-,-,+,+,-,-,0,+,-,-,-,+
V,0,-,0,+,-,+,-,+,0,0,-,0,0,0,+,+,-,-,0,-,-,+,-,-,-,0,-,-,0,-,+
F,0,-,0,+,-,+,-,+,0,0,-,0,0,0,+,+,-,-,0,-,-,+,-,-,-,0,-,-,0,-,-
AY,0,+,-,-,-,+,-,0,+,0,+,-,+,-,-,-,-,-,+,-,-,+,+,-,+,0,+,-,0,-,+
AW,0,+,-,-,-,+,-,0,+,0,+,-,-,-,-,-,-,-,+,-,-,+,+,-,+,0,+,-,0,-,+
AA,0,+,+,-,-,+,-,0,-,0,+,-,0,-,-,-,-,-,+,-,-,+,+,-,-,0,+,-,0,-,+
CH,-,-,0,+,-,-,+,+,0,+,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,+,-,-,0,-,-
UW,0,+,+,-,-,+,-,0,-,0,+,-,0,+,+,-,-,+,-,-,+,+,+,-,-,0,+,-,+,-,+
ER,-,+,-,-,-,+,+,0,-,+,-,-,0,-,-,-,-,-,-,-,-,+,+,-,-,-,+,-,-,-,+
N,+,-,0,+,-,-,+,0,0,-,-,0,0,0,-,-,-,-,0,+,-,+,+,-,-,-,-,-,0,-,+
AH,0,+,+,-,-,+,-,0,-,0,+,-,0,-,-,-,-,-,-,-,-,+,+,-,-,0,+,-,-,-,+
G,0,-,0,+,-,-,-,-,0,0,+,0,0,+,-,-,-,-,-,-,-,+,-,-,-,0,-,-,0,-,+
T,+,-,0,+,-,-,+,-,0,-,-,0,0,0,-,-,-,-,0,-,-,+,-,-,-,-,-,-,0,-,-
VY,0,+,-,-,0,+,-,0,+,0,+,-,+,-,-,-,-,-,+,-,-,+,+,-,+,0,+,-,0,-,+
```

4. Save the file as a .txt.

5. In PCT, go to File > Manage feature systems... > Create feature system from text file

6. Select the file you just created and change only `Column delimiter` to `,` (comma).

APPENDIX E

Lexical statistics by participant

In this appendix, I report the token frequency and type informativity calculated for [a͡ɪ], [ʌ͡i], and [ɔ͡ɪ] for each participant, though participants who were excluded from the study are not included here. Note that lexical statistics for [ɔ͡ɪ] do not vary across participants, as we made the simplifying assumption that speaker lexicons are identical with respect to [ɔ͡ɪ].

This information is also accessible as a .csv file in the supplementary materials at osf.io/4xveb as `summary.csv`. That file also includes lexical measures that I did not use in this dissertation (type frequency and token informativity) as well as information on 95% confidence intervals, overlap of confidence intervals, response curve shape, and more. Unfortunately, though I would have liked to include all of that information here, I am unable to do so for reasons of space and formatting.

Finally, I will mention that nowhere do I include the g, k or x_0 parameters I discuss in §2.2.1. This is because I was unable to reliably estimate these parameters — I was not able to fit a logistic to response curves that do not look fully sigmoidal. The reason for this is simply that a linear response curve corresponds to an infinite number of logistics: a line is described uniquely with only two parameters, while, in order to properly represent both upper and lower asymptotes that are not assumed to be 1 and 0, I must use a four parameter logistic curve. Thus, a unique logistic function cannot be identified for linear responses. An approach that could be taken to compare response curves parametrically would be to fit logistics for curves that suit and calculate k; then for all other curves, calculate slope about $y = 0.5$ — however, these slopes would have potentially quite different interpretations. Also, x_0 for logistics could be compared to $x_{y=.5}$, the value of x for which $y = 0.5$, for all other curves, but this would also not be an apples-to-apples comparison.

participant	a͡ɪ token freq.	ʌ͡i token freq.	ɔ͡ɪ token freq.	ʌ͡i type inform.	a͡ɪ type inform.	ɔ͡ɪ type inform.
p01	79351.2	17639.7	2878	4.392	3.117	4.662
p02	80512.1	16478.8	2878	4.418	3.109	4.662
p03	79673.7	17317.2	2878	4.429	3.108	4.662
p04	79105.2	17885.7	2878	4.402	3.112	4.662
p05	80084.6	16906.3	2878	4.399	3.11	4.662
p07	79606.7	17384.2	2878	4.391	3.112	4.662
p10	79131.6	17859.3	2878	4.418	3.114	4.662
p13	85122.5	11868.4	2878	4.409	3.117	4.662
p14	79734.9	17256	2878	4.41	3.11	4.662
p15	80339.1	16651.8	2878	4.403	3.109	4.662
p16	79665	17325.9	2878	4.433	3.104	4.662
p17	79260.7	17730.2	2878	4.398	3.105	4.662
p18	79322.6	17668.3	2878	4.418	3.108	4.662
p19	80398.9	16592	2878	4.416	3.113	4.662
p21	79367.8	17623.1	2878	4.437	3.099	4.662
p22	84403.7	12587.2	2878	4.429	3.103	4.662
p23	79701.7	17289.2	2878	4.433	3.102	4.662
p24	78959.1	18031.8	2878	4.436	3.105	4.662
p25	79589.6	17401.3	2878	4.422	3.108	4.662

Continued from previous page

participant	a͡ɪ token freq.	ʌ͡i token freq.	ɔ͡ɪ token freq.	ʌ͡i type inform.	a͡ɪ type inform.	ɔ͡ɪ type inform.
p26	79447.4	17543.5	2878	4.417	3.107	4.662
p28	79563.4	17427.5	2878	4.434	3.101	4.662
p29	79711.4	17279.5	2878	4.419	3.111	4.662
p30	80242.6	16748.3	2878	4.431	3.1	4.662
p31	79030.4	17960.5	2878	4.441	3.105	4.662
p32	79248.5	17742.4	2878	4.424	3.102	4.662
p33	79494.8	17496.1	2878	4.427	3.107	4.662
p34	79701.4	17289.5	2878	4.416	3.109	4.662
p41	78784.4	18206.5	2878	4.434	3.105	4.662
p43	79694.1	17296.8	2878	4.408	3.109	4.662
p45	79464.9	17526	2878	4.406	3.108	4.662
p46	79523.4	17467.5	2878	4.411	3.098	4.662

Bibliography

Albright, Adam & Bruce Hayes. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2):119–161. DOI: 10.1016/S0010-0277(03)00146-X.

Allopenna, Paul D, James S Magnuson, & Michael K Tanenhaus. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38(4):419–439. DOI: 10.1006/jmla.1997.2558.

Amano, Shigeaki & Tadahisa Kondo. (2000). *Nihongo-no Goitokusei*. Tokyo: Sanseidō, second ed.

Anderson, Jennifer L, James L Morgan, & Katherine S White. (2003). A statistical basis for speech sound discrimination. *Language and Speech* 46(2–3):155–182. DOI: 10.1177/00238309030460020601.

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

Bermúdez-Otero, Ricardo. (2003). The acquisition of phonological opacity. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds). *Variation within Optimality Theory: Proceedings of the Stockholm Workshop on 'Variation within Optimality Theory'*. Stockholm, 25–36.

Bloomfield, Leonard. (1939). Menomini morphophonemics. *Travaux du Cercle Linguistique de Prague* 8.

Boersma, Paul & Silke Hamann. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25(2):217–270. DOI: 10.1017/S0952675708001474.

Brysbaert, Marc & Boris New. (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. DOI: 10.3758/BRM.41.4.977.

Bybee, Joan. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes1* 10(5):425–455. DOI: 10.1080/01690969508407111.

Chambers, J. K. (1973). Canadian raising. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 18(2):113–135. DOI: 10.1017/S0008413100007350.

Chambers, J. K. (2006). Canadian raising retrospect and prospect. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 51(2-3):105–118. DOI: 10.1017/S000841310000400X.

Clements, George N. (2003). Feature economy in sound systems. *Phonology* 20(3):287–333. DOI: 10.1017/S095267570400003X.

Cohen Priva, Uriel. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6(2):243–278. DOI: 10.1515/lp-2015-0008.

Dailey, Shannon & Elika Bergelson. (2022). Language input to infants of different socioeconomic statuses: A quantitative meta-analysis. *Developmental Science* 25(3):e13,192. DOI: 10.1111/desc.13192.

Davies, Mark. (2012). The Strathy corpus of Canadian English.

de Leeuw, Joshua R, Rebecca A Gilbert, & Björn Luchterhandt. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software* 8(85):5351. DOI: 10.21105/joss.05351.

Fry, Dennis B, Arthur S Abramson, Peter D Eimas, & Alvin M Liberman. (1962). The identification and discrimination of synthetic vowels. *Language and Speech* 5(4):171–189. DOI: 10.1177/002383096200500401.

Gahl, Susanne. (2008). Time and Thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84(3):474–496. DOI: 10.1353/lan.0.0035.

Gahl, Susanne, Yao Yao, & Keith Johnson. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66(4):789–806. DOI: 10.1016/j.jml.2011.11.006.

Ganong, William F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6(1):110–125. DOI: 10.1037/0096-1523.6.1.110.

Gelbart, Ben. (2005). *Perception of foreignness*. Doctoral dissertation, University of Massachusetts Amherst.

Goldsmith, John A. (1995). Phonological theory. In John A Goldsmith (ed). *The Handbook of Phonological Theory*. chap. 1, 1–23, Cambridge, MA: Blackwell.

Hall, Kathleen Currie. (2005). Defining phonological rules over lexical neighbourhoods: Evidence from Canadian raising. In John Alderete, Chung-hye Han & Alexei Kochetov (eds). *Proceedings of the 24th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Press, 191–199.

Hall, Kathleen Currie. (2009). *A probabilistic model of phonological relationships from contrast to allophony*. Doctoral dissertation, Ohio State University.

Hall, Kathleen Currie. (2012). Phonological relationships: A probabilistic model. In Jenny Loughran & Alanah McKillen (eds). *Proceedings from Phonology in the 21st Century: In Honour of Glyne Piggott*.

Hall, Kathleen Currie. (2013). A typology of intermediate phonological relationships. *Linguistic Review* 30(2):215–275. DOI: 10.1515/tlr-2013-0008.

Hall, Kathleen Currie, Blake Allen, Edith Coates, Michael Fry, Serena Huang, Khia Johnson, Roger Lo, Scott Mackie, Stanley Nam, & Michael McAuliffe. (2021). Phonological Corpus Tools 1.5.0.

Hamming, Richard W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal* 29(2):147–160. DOI: 10.1002/j.1538-7305.1950.tb00463.x.

Hart, Betty & Todd R Risley. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H Brookes Publishing Co.

Hay, Jennifer, Janet B Pierrehumbert, & Mary E Beckman. (2004). Speech perception, well-formedness, and the statistics of the lexicon. In John Local, Richard Ogden & Rosaline Temple (eds). *Phonetic interpretation: Papers in laboratory phonology VI.* First ed, chap. 3, 58–74, New York, NY: Cambridge University Press.

Hayes, Bruce. (1995). On what to teach the undergraduates: Some changing orthodoxies in phonological theory. *Linguistics in the morning calm* 3:59–77.

Hazan, Valerie & Sarah Barrett. (2000). The development of phonemic categorization in children aged 6-12. *Journal of Phonetics* 28(4):377–396. DOI: 10.1006/jpho.2000.0121.

Horst, Jessica S & Michael C Hout. (2016). The Novel Object and Unusual Name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods* 48(4):1393–1409. DOI: 10.3758/S13428-015-0647-3.

Hussain, Qandeel & Shigeko Shinohara. (2019). Partial devoicing of voiced geminate stops in Tokyo Japanese. *Journal of the Acoustical Society of America* 145(1):149–163. DOI: 10.1121/1.5078605.

Idsardi, William J. (2006). Canadian raising, opacity, and rephonemicization. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 51(2-3):119–126. DOI: 10.1017/S0008413100004011.

Iskarous, Khalil, Christine Mooshammer, Philip Hoole, Daniel Recasens, Christine H Shadle, Elliot Saltzman, & Douglas H Whalen. (2013). The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *Journal of the Acoustical Society of America* 134(2):1271–1282. DOI: 10.1121/1.4812855.

Itô, Junko & Armin Mester. (1993). Japanese phonology constraint domains and structure preservation. In John Goldsmith (ed). *The Handbook of Phonological Theory*. First ed, chap. 29, 817–838, New York, NY: Blackwell.

Itô, Junko & Armin Mester. (2017). The phonological lexicon. In Natsuko Tsujimura (ed). *The Handbook of Japanese Linguistics*. First ed, chap. 3, 62–100, Malden, MA: Blackwell Publishers.

Jongman, Allard, Zhen Qin, Jie Zhang, & Joan A Sereno. (2017). Just noticeable differences for pitch direction, height, and slope for Mandarin and English listeners. *The Journal of the Acoustical Society of America* 142(2):EL163–EL169. DOI: 10.1121/1.4995526.

Joos, Martin. (1942). A phonological dilemma in Canadian English. *Language* 18(2):141–144. DOI: 10.2307/408979.

Jun, Sun-Ah & Jihyeon Cha. (2015). High-toned [il] in Korean: Phonetics, intonational phonology, and sound change. *Journal of Phonetics* 51:93–108. DOI: 10.1016/j.wocn.2015.05.002.

Kager, René. (2008). Lexical irregularity and the typology of contrast. In Kristin Hanson & Sharon Inkelas (eds). *The nature of the word: Studies in honor of Paul Kiparsky*. 397–432, Cambridge, MA: MIT Press.

Kawahara, Shigeto. (2005). Voicing and geminacy in Japanese: An acoustic and perceptual study. *University of Massachusetts Occasional Papers in Linguistics* 31:87–120.

Kingston, John, Joshua Levy, Amanda Rysling, & Adrian Staub. (2016). Eye movement evidence for an immediate Ganong effect. *Journal of Experimental Psychology: Human Perception and Performance* 42(12):1969–1988. DOI: 10.1037/xhp0000269.

Kreiman, Jody, Bruce R Gerratt, & Sameer ud Dowla Khan. (2010). Effects of native language on perception of voice quality. *Journal of Phonetics* 38(4):588–593. DOI: 10.1016/j.wocn.2010.08.004.

Ladd, D Robert. (2006). "Distinctive phones" in surface representation. In Louis Goldstein, Douglas H Whalen & Catherine T Best (eds). *Laboratory Phonology* 8. First ed, chap. 1, 3–26, Berlin: Walter de Gruyter GmbH.

Liberman, Alvin M, Katherine S Harris, Howard S Hoffman, & Griffith C Belver. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54(5):358–368. DOI: 10.1037/h0044417.

Lohman, Arne. (2018). Time and thyme are NOT homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis. *Language* 94(2):e180–e190. DOI: 10.1353/lan.2018.0032.

Luce, Paul A, Conor T McLennan, & Jan Chance-Luce. (2003). Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In Jeffrey S Bowers & Chad J Marsolek (eds). *Rethinking implicit memory*. chap. 9, 197–214, Oxford, UK: Oxford University Press.

Maddieson, Ian. (1985). Borrowed sounds. In *UCLA Working Papers in Phonetics 61*. UCLA, 51–64.

Maye, Jessica & LouAnn Gerken. (2000). Learning phonemes without minimal pairs. In S Catherine Howell, Sarah A Fish & Thea Keith-Lucas (eds). *Proceedings of the 24th annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, 522–533.

Mayer, Connor. (2020). An algorithm for learning phonological classes from distributional similarity. *Phonology* 37(1):91–131. DOI: 10.1017/S0952675720000056.

McMurray, Bob, Ani Danelz, Hannah Rigler, & Michael Seedorff. (2018). Speech categorization develops slowly through adolescence. *Developmental Psychology* 54(8):1472–1491. DOI: 10.1037/dev0000542.

Morimoto, Maho. (2020). *Geminated liquids in Japanese: A production study*. Doctoral dissertation, University of California, Santa Cruz.

Nixon, Jessie S, Jacolien Van Rij, Peggy Mok, Harald R Baayen, & Yiya Chen. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language* 90:103–125. DOI: 10.1016/j.jml.2016.03.005.

Pater, Joe & Elliott Moreton. (2014). Structurally biased phonology: Complexity in learning and typology. *The EFL Journal* 3(2):1–44.

Pedersen, Eric J, David L Miller, Gavin L Simpson, & Noam Ross. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ* 7:e6876. DOI: 10.7717/peerj.6876.

Peperkamp, Sharon, Michèle Pettinato, & Emmanuel Dupoux. (2003). Allophonic variation and the acquisition of phoneme categories. In Barbara Beachley, Amanda Brown & Frances Conlin (eds). *Proceedings of the 17th Annual Boston University Conference on Language Development* 2. Somerville, MA: Cascadilla Press, 650–661.

Reinisch, Eva & Matthias J Sjerps. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics* 41(2):101–116. DOI: 10.1016/j.wocn.2013.01.002.

Scobbie, James M. (2005). The phonetics phonology overlap. *QMUC Speech Science Research Centre Working Paper* 1:1–30.

Scobbie, James M & Jane Stuart-Smith. (2008). Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English. In Peter Avery, B Elan Dresher & Keren Rice (eds). *Contrast in Phonology: Theory, Perception, Acquisition*. 87–114, New York, NY: Mouton de Gruyter.

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27(3):379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Slawinski, Elzbieta B & Laurie K Fitzgerald. (1998). Perceptual development of the categorization of the /ɹ-w/contrast in normal children. *Journal of Phonetics* 26(1):27–43. doi: 10.1006/JPHO.1997.0057.

Sperry, Douglas E, Linda L Sperry, & Peggy J Miller. (2019). Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development* 90(4):1303–1318. doi: 10.1111/cdev.13072.

SR Research Ltd. (2020). SR Research Experiment Builder 2.3.38.

SR Research Ltd. (2021). EyeLink Data Viewer 4.2.1.

Steffman, Jeremy. (2020). *Prosodic prominence in vowel perception and spoken language processing*. Doctoral dissertation, University of California, Los Angeles.

Steffman, Jeremy & Megha Sundara. (2024). Disentangling the role of biphone probability from neighborhood density in the perception of nonwords. *Language and Speech* 67(1):166–202. doi: 10.1177/00238309231164982.

Tamaoka, Katsuo & Shogo Makioka. (2004). Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments, & Computers* 36(3):531–547. doi: 10.3758/BF03195600.

Thiessen, Erik D & Philip I Pavlik Jr. (2016). Modeling the role of distributional information in children's use of phonemic contrasts. *Journal of Memory and Language* 88:117–132. doi: 10.1016/j.jml.2016.01.003.

Van Rij, Jacolien, Martijn Wieling, Harald R Baayen, & Heddrick van Rijn. (2022). itsadug: Interpreting time series and autocorrelated data using GAMMs.

Vance, Timothy J. (1987). "Canadian Raising" in some dialects of the Northern United States. *American Speech* 62(3):195. DOI: 10.2307/454805.

Vitevitch, Michael S & Paul A Luce. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers* 36(3):481–487.

Vitevitch, Michael S & Paul A Luce. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics* 2(1):75–94. DOI: 10.1146/annurev-linguistics-030514-124832.

Watson, Janet C E. (2002). *The phonology and morphology of Arabic*. New York, NY: Oxford University Press, first ed.

Werker, Janet F, H Henry Yeung, & Katherine A Yoshida. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science* 21(4):221–226. DOI: 10.1177/0963721412449459.

Wickham, Hadley. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer Cham, first ed.

Winn, Matthew B. (2019). Make formant continuum.

Wood, Simon N. (2017). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC, second ed.