UNIVERSITY OF CALIFORNIA SAN DIEGO

Functional pressures and linguistic typology

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Linguistics

by

Eric Meinhardt

Committee in charge:

Professor Eric Baković, Co-Chair
Professor Leon Bergen, Co-Chair
Professor Richard Futrell
Professor Marc Garellek
Professor Jason Schweinsberg

2021

The dissertation of Eric Meinhardt is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am deeply indebted to my (past and present) co-advisors Eric Baković, Leon Bergen, and Andy Kehler for their support and advice throughout my PhD. Feedback from other committee members (past and present) Marc Garellek, Richard Futrell, Jason Schweinsberg, and Massimo Vergassola was also important to shaping writing, the direction of my research, or helping me to resolve thorny technical questions. Eric, Andy, and Farrell Ackerman deserve special thanks for helping me make it through the first few years of graduate school, for helping me appreciate what linguistics uniquely brings to the language sciences, and for both encouraging my wide-ranging interests and (more importantly) helping me learn to focus them. Much of the dissertation and exam work or coursework that preceded it could not have been done without Marc's instruction, advice, and reading suggestions on phonetics. At the other end of my PhD, Leon's advice and guidance was crucial to this dissertation's direction and making it feasible to finish.

I am also grateful to Eric, Leon, Farrell, Rob Malouf, Anna Mai, Adam McCollum, Nadia Polikarpova, Shraddha Barke, and Rose Kunkel as thoughtful collaborators without whose skills and perspectives the interdisciplinary research I have worked with them on would not have been possible. As sources of support throughout the ups and downs of a PhD, as sounding boards, and as intellectual partners-in-crime, I am very thankful for friends both inside the Linguistics Department — Anna, Adam, Kati Hout — and out — Jack Berkowitz, Alex Kuczala, Julieta Gruszko, Shauna Kravec, and Ben Kellman.

I would also like to express my appreciation for Linguistics staff and Social Sciences Computing Facility (SSCF) staff, especially Alycia Randol and Silas Horton.

Last but not least, I would like to acknowledge those who encouraged and inspired me to pursue research in some blend of linguistics, computer science, and cognitive science: my parents, my high school German teacher Mark Wagner, and University of Rochester

professors Chris M. Brown, Robbie Jacobs, Scott Paauw, and Lenhart Schubert. The final chapter of this dissertation would not have been possible without the patient introduction to theoretical evolutionary biology, mathematical population genetics, and evolutionary game theory that University of Rochester professors H. Allen Orr, Dan Garrigan, and Paulo Barelli offered during my 'Take Five' scholarship year.

The financial support of the UC San Diego Division of Social Sciences and Academic Senate were key to allowing time to work on this dissertation, and a Titan V GPU donated by the NVIDIA Corporation was very helpful for accelerating computations vital to two of the chapters.

Figures 2.2a-b are Figures 1-2 from Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995), and are reprinted with permission from AAAS.

Chapter 3 was coauthored with Eric Baković and Leon Bergen, and is very similar to the submitted manuscript that has since been edited and published in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1991-2002). The dissertation author was the primary investigator and author of this paper.

Chapter 4 was coauthored with Farrell Ackerman and Robert Malouf, and is very similar to the submitted manuscript that has since been edited and will be published in the Cambridge University Press volume *Morphological Typology and Linguistic Cognition*. The dissertation author was the primary investigator and author of this paper.

VITA

2013	B. A. in Brain & Cognitive Sciences and Linguistics, University of Rochester
2021	Ph. D. in Linguistics, University of California San Diego

ABSTRACT OF THE DISSERTATION

Functional pressures and linguistic typology

by

Eric Meinhardt

Doctor of Philosophy in Linguistics

University of California San Diego, 2021

Professor Eric Baković, Co-Chair

Professor Leon Bergen, Co-Chair

The explanation of linguistic variation and change is one of the central questions in the language sciences. *Functional* explanations focus on how the needs and abilities of language users shape the distribution of linguistic structures that typically conventionalize — e.g. structures that are harder to perceive or learn accurately are less likely to conventionalize accurately. *Perceptibility effects* are common sound patterns that seem closely related to the relative confusability of different speech sound sequences. One class of explanations — phonological accounts — have assumed speakers (implicitly) know how confusability varies as a function of immediately adjacent sounds, and that this is a rich enough description of confusability to explain much of perceptibility effects. Chapter 2 shows that the perceptibility of tokens of any given sound in American English systemati-

cally varies based on a listener's incrementally-adjusted expectations about what the speaker intends to say, and shows that this variation is significantly greater than variation due to immediately adjacent sounds. To derive this result, I present a computational psycholinguistic model of word recognition and apply it to experimental confusability data and a transcribed lexicon of 10^4 words. I conclude that phonological accounts of perceptibility effects need to be much more complicated and less modular than currently appreciated, and are consequently less plausible. Chapter 3 applies the same word recognition model and novel information-theoretic measures of confusability to two conversational corpora and shows that words that are more contextually confusable are lengthened in contexts where they are more confusable, and shortened where they are less so. This is a crucial step towards a linking hypothesis between the realtime perceptibility of different speech sound sequences and conventionalized perceptibility effects. Chapter 4 considers morphology. Prior research has observed an inverse relation between morphological complexity and demographic variables like speech community size and proportion of adult learners. Recent work has hypothesized that higher complexity may be helpful to child learners, and that populations with differing demographics constitute environments with different 'selection pressures' for language variants to 'evolve' in. I argue that mathematical formulations of Darwinian evolution suggest a more likely explanation: 'neutral' change caused by random fluctuations in variant frequency ('drift') is much more powerful in small populations and can easily overwhelm selection relative to large populations.

Chapter 1

Introduction

The following chapters of this dissertation are three case studies of the role that pressures for communication and learning have in explaining the diversity of linguistic structures linguists have observed and their relation to linguistic theory. In all three cases, I focus on patterns in the structure of words and speech sounds.

Chapters 2 and 3 concern *perceptibility effects* — cross-linguistically common patterns in speech sounds and sound changes that are closely related to patterns in what speech sounds are easier or harder to perceive in the context of which other sounds. There are two classes of explanation for perceptibility effects, which I will term *diachronic-phonetic* accounts and *synchronic-phonological* for brevity. The main explanation of the first class is that listener-learners mishear sounds that are confusable and/or misattribute what they hear to being part of the grammar of the language (J. J. Ohala, 1993). A variant of this *listener-error* account also hypothesizes a role for the *choices of speakers*. It suggests that how speakers pronounce words (and therefore what listener-learners are exposed to, and shaping what kinds of errors they are likely to make) is shaped by their communicative goals: speakers selectively enhance aspects of pronunciation that make listeners more likely

to understand what the speaker means, and underarticulate aspects that likely won't hurt understanding (Lindblom, 1990).

A second broad class of explanation argues that the prevalence of perceptibility effects is also partly explained by grammatical knowledge that *directly references* facts about the relative confusability of different speech sounds in different local environments (Steriade, 2001b): instead of perceptibility effects in sound changes and grammatical sound patterns being an indirect, eventual outcome of conventionalization initially caused by variation in naturalistic listening and speech of particular languages, perceptibility effects are explained by grammatical knowledge that directly references a speaker-internal model of the relative perceptibility of different speech sounds in the context of immediately adjacent speech sounds. A subset of research in this class of explanations has further proposed that this knowledge might be *biologically innate*, and that this contributes to explaining the prevalence of perceptibility effects (Wilson, 2006).

The second chapter of this dissertation — documenting research conducted with the aid and advice of Eric Baković, Leon Bergen, Marc Garellek, and Andrew Kehler — argues against synchronic-phonological explanations of perceptibility effects. In this chapter, I claim that these accounts are based on an unintentional and incorrect assumption about perception and how it plays out in speech processing, namely the assumption that the scope of variation in perceptibility of speech sounds caused by context is limited to the relatively *local* phonotactic context that laboratory studies have examined. I show this is incompatible with what is known about perception generally and psycholinguistics specifically, and through a mathematically explicit model of word recognition applied to data on the confusability of American English speech sounds (Warner, McQueen, & Cutler, 2014), show that perceptibility is, in fact, sensitive to a much more expansive notion of context, and that the perceptibility of a given speech sound varies both more and differently than this second class of explanations assumed. I argue that revising this type explanation to

accurately reflect facts of perceptibility is most consistent with the class's stated motivations, and yet also renders it psychologically much less plausible as a result of the changes in the architecture of phonological knowledge it entails. I conclude that the weight of evidence supports diachronic-phonetic explanations of perceptibility effects instead.

The third chapter — a conference submission written jointly with Leon Bergen and Eric Baković — tests a key prediction of the 'speaker-choice' variant of diachronic-phonetic explanations. Using a variant of the model developed in the first chapter applied to corpora of conversational American English speech (Calhoun et al., 2010; Pitt et al., 2007), the chapter examines whether speakers typically lengthen content words (making them easier to understand) when they are more contextually confusable and typically shorten them (economizing on production cost) when they are less contextually confusable, and finds that they do.

The final chapter — a book chapter written with Rob Malouf and Farrell Ackerman — considers morphology rather than phonetics and phonology, and ultimately argues against a functionally-oriented explanation of typology in favor of population-level explanations. The *enumerative complexity* of a language's morphology concerns details like the number of morphosyntactic categories, the number and variation of formatives used to encode them, and the combinatorics of how those formatives appear in the language (Ackerman & Malouf, 2013; Stump & Finkel, 2013). Linguists have observed that *high* enumerative complexity in morphology seems to be uncommon among languages spoken by historically *exoteric* communities (Wray & Grace, 2007) — ones with larger populations of speakers, covering a large area, with lots of language contact, and often with notably large proportions of adult second-language learners at some point in their history — and that instead higher complexity seems to be found principally in languages spoken by communities historically lacking these qualities or displaying the opposite trends — *esoteric* communities. Psycholinguists, meanwhile, have found some empirical evidence for the idea that adults seem to have greater

difficulty than children in learning (enumeratively) complex morphology — in other words, that forms and languages with lots of forms that are less enumeratively complex are more easily learned by — *well-adapted* for learning by — adult second-language learners.

Lupyan and Dale (2010, 2015, 2016a) and Dale and Lupyan (2012) speculate that high enumerative complexity may be adaptive for the learning mechanisms available to children, and that this explains *why* it is associated with esoteric populations. They couch this hypothesis about morphological typology in terms of *cultural evolution*: adult second-language learners (present in exoteric populations) select against high complexity and for low complexity, and child learners select for high complexity. The differential net pressures in these different types of populations, they hypothesize, lead to differential survival and propagation of certain types of forms and systems of forms in those different types of populations.

The fourth chapter argues that, in light of the lack of empirical evidence or any detailed model of child learning that explicates Lupyan and Dale's hypothesis about high complexity supporting child learning, evolutionary theory and mathematical models of evolutionary processes suggest other hypotheses are *a priori* simpler and much more likely to explain observed trends in esoteric populations and morphological complexity. While Lupyan and Dale (2010) are correct that language and culture can be understood as Darwinian evolutionary systems comparable to biological evolution in a well-defined, abstract sense, the fourth chapter argues that they have missed several of the most important lessons of 20th century evolutionary theory — ones that apply to any Darwinian evolutionary system. As noted in the chapter, debate over the relative likelihood and explanatory burden of random changes vs. of adaptive changes in explaining observed patterns in evolution has been a key part of modern evolutionary theory and motivated both rigorous mathematical theory development and a high burden of evidence for adaptive explanations (Gould & Lewontin, 1979; Pigliucci & Kaplan, 2000; Stephens, 2008). As elaborated in the chapter,

such mathematical work was instrumental in determining that the evolution of smaller populations is much more sensitive to random fluctuations than that of large ones, and that in small populations the effects of even relatively strong adaptive pressures can be overwhelmed by such random fluctuations, meaning that even traits that are selected *against* and that would disappear in large populations can arise and stably persist in small populations. Turning to work specifically on language, the chapter also reviews recent work on language change and finds it consistent with these observations as well. Consequently, if a linguistic pattern appears to be generally absent in large populations, but when present, is present principally in populations that are small, then (*ceteris paribus*) it is more likely to be *neutral* than adaptive, and could even plausibly be *maladaptive*. Lacking forthcoming specific empirical evidence for or a well-motivated mathematical model linking high morphological complexity to ease of learning in children (but crucially not adults), the chapter concludes that Lupyan and Dale's hypothesis about the typology of esoteric speech communities is less likely than the assumption that high enumerative complexity has no specific benefit to language learning in children.

Chapter 2

Perceptibility Effects and Constraint-Based Phonological Theory

2.1 Introduction

The field of phonetics is principally concerned with realtime, continuous, gradient, physical, measurable, and contextually often highly variable properties of speech sounds – their production, acoustics, and perception. Phonology, in contrast, is concerned with the productive generalizations speakers of a language have about sound patterns in their language, where the notion of speech sound is discrete, abstract and categorical, and patterns include things like which sounds tend to (or must or cannot) occur next to which other sounds. For example, it is knowledge of phonological generalizations about English sound patterns that allows a native speaker to conclude that even though neither /blɪk/ (*blick*) nor /bnɪk/ (*bnick*) are actual, meaningful words of English and even though they differ only in one speech sound, /blɪk/ is a plausible English word in a way that /bnɪk/ isn't: English-like

words do not and (at least currently) *could* not start with sound sequences like /bn/.[1] While both phonetics and phonology study and characterize knowledge that speakers have of sounds in their language, the nature of that knowledge is quite different. The phenomena I examine in this chapter concern both fields.

In particular, work in phonetics has established that the confusability of speech sounds varies across different natural classes,[2] and can be strongly affected by adjacent speech sounds. For example, the contrast between [b] and [p] is easier to hear between two vowels than before a stop:[3] [aba] and [apa] are easier to discriminate than [abta] and [apta] (Wright, 2004). Cross-linguistically, natural languages tend to avoid distinguishing words using contrasts in phonotactic contexts where the contrast in question is difficult to perceive (Steriade, 2001b). Similarly, patterns of sound change – both which are well-attested and which are rare – seem to correspond to laboratory-observable patterns of confusability (Blevins, 2008; Garrett & Johnson, 2011; Hansson, 2008; J. J. Ohala, 1981). These patterns in phonology and historical linguistics constitute *perceptibility effects*. They are the empirical phenomenon considered in this chapter. The broad scientific question considered in this chapter is what I will refer to as **the Mechanism Question**:

(2.1.1) What are the mechanisms linking gradient phonetic facts about realtime perceptibility to categorical patterns in the phonology of individuals and language change in populations?

To contextualize exactly what the contribution of this chapter is, I will group

[1]Instead, speakers typically 'repair' words similar to /bnɪk/ (e.g. the Nordic surname *Knuth*) by inserting a vowel between the two consonants, producing [bənɪk] (*buhnick*).

[2]A natural class of speech sounds are sounds in a language that all share notable articulatory and/or acoustic properties, and that typically are treated similarly in phonological patterns. *Vowels* and *consonants* are two familiar (if very broad) examples of natural classes.

[3]Stops (e.g. /b/, /p/, /t/, /d/, /k/, /g/) are a natural class of consonants that are all produced by forming a complete stoppage of airflow from the lungs, causing a build up of pressure behind that stoppage, and then suddenly releasing that pressure, causing a transient burst of noise.

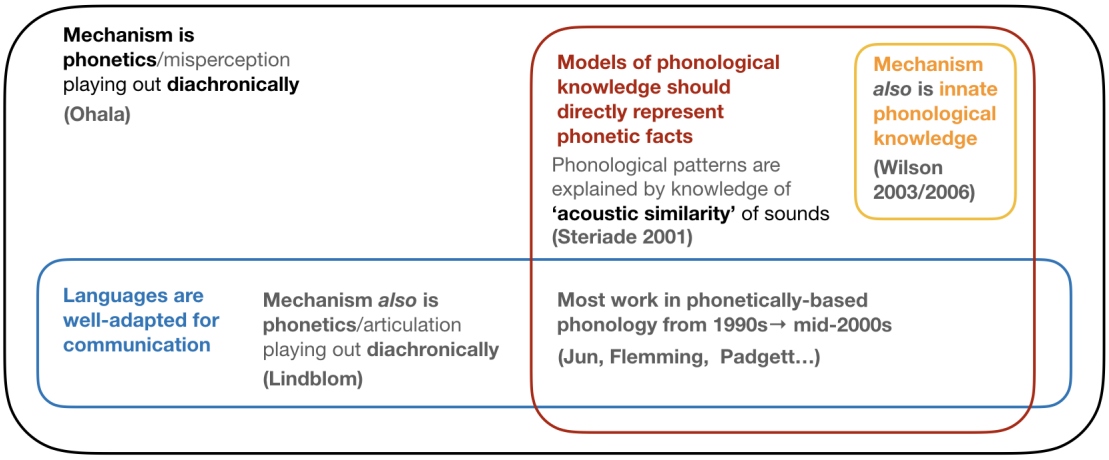


Figure 2.1: Classification of explanations for perceptibility effects

existing explanations for perceptibility effects per Figure 2.1. It is uncontroversial that contrasts which are difficult for learners to accurately perceive are more likely to be misperceived (J. J. Ohala, 1993) and that this is a key part of explaining perceptibility effects, but progress in elucidating more precise theories about the mechanisms linking individual misperception episodes to changes in the beliefs about categorical phonological patterns in individuals or many such changes across populations and over time has been elusive. As well, a large body of work in phonology (Hayes & Steriade, 2004; Hume & Johnson, 2001) has introduced formal theories of grammar where the relative confusability of different sounds is *directly* represented as part of grammatical knowledge. However, because the mission of (spoken-language) phonology is generally taken to be modeling the productive knowledge of *speakers* about sound patterns rather than the knowledge of *analysts*, the causal role and psychological reality of such theories (i.e. how strongly phonological theories relate to the Mechanism Question) remained unclear at best and somewhere between unconvincing and redundant with existing phonetic explanations at worst (Zhang & Lai, 2006). In the early-to-mid 2000s (Wilson, 2003, 2006; Zhang & Lai, 2006; Zuraw, 2007),

however, research on phonological accounts began to explore the hypothesis that an *innate* (and explicitly phonological) *inductive bias* for learning more perceptible phonological patterns may contribute to explaining the prevalence of perceptibility effects. Finally, cross-cutting phonetic (see e.g. J. Kirby, 2013; Lindblom, Guion, Hura, Moon, & Willerman, 1995) and phonological accounts (e.g. Flemming, 2001a) are theories hypothesizing that perceptibility effects reflect *communicative pressures* on the structure of natural language, as well as both phonetic (e.g. J. J. Ohala, 1997) and phonological (e.g. Steriade, 2001b) accounts of perceptibility effects that are highly critical or less committed (respectively) to such an explanation for perceptibility effects.

In this chapter, I will argue that

(2.1.2)
a. Phonological accounts are based on an incomplete understanding of perception that assumes relative perceptibility of any given speech sound in any given phonotactic context is far less variable across and even within languages than it actually is.
b. They make incorrect predictions about perceptibility effects as a result.
c. Their stated motivations for incorporating phonetic facts into grammatical representations are most consistent with having a psychologically accurate model of perceptibility, but revising phonological accounts to accurately reflect systematic variation in perceptibility requires a dramatically different, more complex, and far less plausibly innate model of phonological knowledge than previously appreciated.

To make this argument, I first review what is known about the psycholinguistics of *contextual perceptibility* of words and the speech sounds within them, showing that while phonological accounts acknowledge and make use of the effect of local *phonotactic* context on confusability, they have neglected the known effects of *epistemic* context: listeners

combine bottom-up acoustic data about the currently unfolding segment with incrementally updated top-down expectations about what they are perceiving (Marslen-Wilson & Tyler, 1980; Norris & McQueen, 2008). This means that calculation of the confusability of any given speech sound token requires consideration of the gradient structure of the *entire lexicon* as well as any other stable context cues (linguistic or not).

To show the scope and magnitude of empirical differences between out-of-context and in-context perceptibility, I derive, implement, and analyze a computational model of word recognition simple enough to analyze and relate to existing psychoacoustic data (Warner et al., 2014) and complex enough to model bottom-up acoustic confusability and a stable source of top-down, incrementally updating expectations (the lexicon). The structure of the model allows for a precise, compact description of the incorrect predictions existing phonological accounts should make, and its implementation using real psychoacoustic data on perceptibility permits approximate measurement of the difference between out-of-context vs. in-context confusability of the phonological inventory of American English. The results clearly demonstrate that for most speech sounds, these differences are significant.

I conclude by discussing why this means existing phonological accounts cannot accomplish their stated aims as they are currently formulated, and that revising them to accurately model confusability would render them even more implausible than presently appreciated, particularly contemporary reformulations positing an innate bias and offering the otherwise most compelling case for the relevance of a direct representation of phonetic facts in phonological knowledge.

2.2 Background

In this section, after briefly reviewing the differences between phonetics and phonology, I introduce the phenomenon to be investigated and the two main classes of existing explanations in more detail.

2.2.1 Phonetics vs. phonology

This subsection briefly introduces and differentiates phonetics and phonology from each other as fields, reviews some vocabulary and notation, and the difference in character between phonetic and phonological explanations.

Though the border between the two is sometimes not clear,[4] the field of *phonetics* conventionally centers on the physical properties of speech sounds and events relatively immediately adjacent to them — the articulation of speech sounds, their measurable acoustic properties, and their perception, without necessarily a specific focus on the role or variation of any of these things in a specific language relative to others. For example, the /p/ sound in English is an instance of a class of sounds called *stops* or *plosives*; it is produced by forming a complete stoppage of airflow from the lungs, causing a build up of pressure behind that stoppage as the lungs continue to contract, and then suddenly releasing that, causing a sudden drop in pressure and transient burst of noise. In the articulation and acoustics of the /p/ sound in *pin*, *spin*, and *stop*, /p/ is produced slightly differently: the realization of /p/ in *pin* is produced with a puff of air ('aspiration'), but this is not present in *spin*, and in the case of *stop*, you may not even (immediately or audibly) open your mouth and release the build up of pressure. In broad strokes, phonetics considers each of these a different object and a phonetic *transcription* of the typical production of these words would reflect this as [pʰɪn], [spɪn], and [stɑp˺].

Phonology is conventionally concerned with productive grammatical knowledge that individuals have about *sound patterns* in the languages they speak, where the notion of 'speech sound' is more abstract, discretizable, categorical, less contextually variable, often at least somewhat arbitrary, and centrally defined and examined with reference to other aspects of language structure than it typically is in phonetics. Examples of questions

[4] Or even argued to not exist — see e.g. J. J. Ohala (1990b).

phonology is concerned with include the inventory of abstract sound units that are used to build larger units of form, the restrictions on what sounds can be next to each other ('phonotactics'), and how sounds change in different contexts ('alternations'). With respect to phonology, what is interesting about the different manifestations of the sound /p/ in the three words in the example above is that they are *predictable variants* of what is effectively the *same object* – a unit of form called a *phoneme* – in the larger context of English sound patterns, meaning that a native speaker would identify them as the same sound, and that a phonologist could plausibly transcribe these three words as [pɪn], [spɪn], and [stɑp].[5] That is, on the one hand, there are no words (or other kinds of units of meaning) that are distinguished by aspiration in English, unlike in other languages (e.g. Hindi). On the other hand, the change between variant realizations of the phoneme /p/ in *pin*, *spin*, and *stop* is not an arbitrary property of those words – the particular realization of /p/ is predictable across other words with reference to the position of /p/ within that word. In fact, this pattern of variation holds not just of /p/, but of other stop sounds in English as well – it is a *productive* pattern that English speakers apply to items they may never have encountered before.

To offer a second example of phonological patterns and knowledge, consider variation in the English plural. For many words, this is an [s]: the plurals of /bip/ (*beep*), /but/ (*boot*), and /bʊk/ (*book*) are [bips], [buts], and [bʊks]. This is not always the case, though: the plurals of /bɑrb/ (*barb*), /bʌd/ (*bud*), and /bʌg/ (*bug*), among others, have a [z] instead. The generalization relating these two cases rests on how [s] and [z] differ from each other in the same way that [p] and [b], [t] and [d], and [k] and [g] do. A consonant can be classified by specifying three properties of the speech organs involved in its articulation: the *place* in the vocal tract where these articulators are acting, the *manner* in which they are acting, and *voicing*, or whether or not the vocal folds are vibrating during articulation. The

[5]Note that *square* brackets refer to phonetic realization where *slashes* indicate a more abstract phonological representation.

two sounds in each of the pairs mentioned are the same across the first two dimensions and differ only in the third. The generalization about the English plural in the data offered, then, is that the realization (or 'output') of the underlying (or 'input') /s/ changes to match (or 'assimilates to') the voicing of the final speech segment in the word the plural is attaching to. The fact that English speakers productively and consistently generalize this pattern to novel words that they've never heard before (and couldn't therefore be simply memorizing) is further evidence that this is a productive pattern that is part of what it is to know the sound patterns of English as a native speaker: if /wʌg/ (*wug*) or /blɪk/ (*blick*) were actual words of English, their plurals would be [wʌgz] and [blɪks], rather than *[wʌgs] or *[blɪkz].[6]

As the next section details, this chapter is concerned with phenomena at the intersection of phonetics and phonology, and existing work is essentially arguing about how and why the two seem to be closely related. In the broader context of linguistic and phonological theory, however, this is quite unusual: phonetics and phonology have traditionally been considered to be relatively distinct aspects of what it is to know how to speak and comprehend a language. In a nutshell, the orthodox position in linguistics (so-called *generative linguistics*) since the mid-20th century (Chomsky, 1957, 1965; Chomsky & Halle, 1968) is that

(2.2.3) The object of a scientific description of a natural language is to treat it as a *formal language*, and to therefore offer a description of *all and only those strings that are part of that language*.

In the context of sentences of natural language especially, the infinite number of structures that are possible in a given human language means that as an analyst, simply listing them is inadequate. Similarly, from a psychological perspective, the fact that children learn languages relatively effortlessly and that both children and adults regularly produce sentences

[6]The * to the left of a transcribed word is a notational convention indicating that it is ill-formed.

(as well as words) that they've never encountered (and generally consistently as others do in their language community) means they are doing more than memorization – they infer generalizations that go beyond the narrow content of what they've observed. By treating natural languages as formal languages, however, we can characterize knowledge of a language in terms of

(2.2.4) A discrete, abstract, *computational* procedure – a finite set of rules for manipulating symbols (viz. a *grammar*) – that can *generate* all and only those strings in the language.

This means that knowledge of phonological patterns like the English plural have been traditionally taken to be characterizable by a function (traditionally, a string rewriting rule) mapping from the concatenation of the *underlying forms* for a word stem (e.g. /bʌd/) and the English plural /s/ to their combined surface form ([bʌdz]). The content of a phonological theory of a language (within the tradition of generative linguistics) is taken to be the description of those functions, how they are composed, and the representational content of the objects those functions are defined on.

As you may imagine, it is not obvious how continuous and gradient facts of articulation and perception fit into this picture of linguistic knowledge. To relate (or further distinguish) the two, generative linguistics introduced the *competence-performance distinction* (Chomsky, 1965): the idealized, abstracted notion of what it is to know a language described above is *competence*, a model of the *capacity* of a speaker to grasp and know things about their language. *Performance*, in contrast, involves actually *doing* things (e.g. producing specific utterances) with that knowledge, and may differ from competence for reasons that aren't obviously intimately or specifically linguistic, like working memory limitations or being distracted. Phonology, as grammatical knowledge, has always been recognized as part of competence, whereas phonetics has at least traditionally been considered

strongly part of performance.

In sum then, the character of phonetic and phonological knowledge (particularly as modeled by generative linguistics) overlap, but are distinct from one another and concern different aspects of speech sounds and what it is to know how to speak and perceive a given natural language.

2.2.2 Perceptibility effects in phonotactics

Experimental work in phonetics and laboratory phonology has found strong evidence that some classes of speech sounds (e.g. stops) are much more likely to be correctly perceived by listeners when they are in some structural environments compared to others (e.g. before vowels vs. all other environments), typically as a result of the relative strength of specific *transitional acoustic cues* compared to others and a baseline of relatively weak *internal acoustic cues* during the production of at least one of a pair of adjacent speech sounds (Jun, 2004; Wright, 2004, and citations within each).

For example, strong acoustic cues to the place of articulation of oral stops (e.g. whether an oral stop is /p/ vs. /t/ vs. /k/) are in the release of the stop and how it affects the initial acoustics of the following sound, as in the transition from an oral stop to a subsequent sound like a vowel whose steady-state internal acoustic cues are relatively strong, regular, and periodic (Malécot 1958, Wright 2004, §2.1.3). This means that if one stop is immediately followed by another stop, the first stop is denied an environment for a release and good cues as to its identity. Place contrasts between nasals (e.g. distinguishing /n/ vs. /m/) are similar (Malécot, 1956) and are relatively weak to begin with (J. J. Ohala, 1990a). When they are the first part of a consonant cluster[7] and consequently unreleased, then, the contrast between different nasals is especially hard to hear, and as a result the acoustic cues to the place of articulation of the *following* sound dominate. Concretely: it's

[7] A sequence of consecutive consonants.

easier to distinguish [apna] from [apma] than [anpa] from [ampa], and you are more likely to misperceive [anpa] as [ampa] than the reverse (J. J. Ohala, 1990a). I.e., the perceived difference between [n] and [m] is greater in the context [voiceless obstruent__vowel] than in the context [vowel__voiceless obstruent]. To compactly represent such statements, I will make use of notation from Steriade (2001b): if

(2.2.5) Δ(n,m)

denotes the perceptible difference between segment sequences [n] and [m], then

(2.2.6) Δ(n/p_a, m/p_a) >> Δ(n/a_p, m/a_p)

compactly expresses the statement about relative perceptibility of nasals in different phonotactic environments.

(2.2.7) [n-m]/p_a >> [n-m]/a_p

is glossable as 'the contrast between [n] and [m] in the environment p_a is greater than in the environment a_p', and does the same even more efficiently.

Observations like this are about online, gradient differences in how easily speech sounds can be identified and discriminated (i.e. 'phonetic' observations about acoustics), but such effects have a close relationship with well-attested conventionalized, categorical patterns (i.e. 'phonological' observations) across diverse languages without any historical relationship, as well as with common sound changes across a variety of languages (Blevins, 2008). Consider the example above, for instance: nasals and unreleased oral stops (as when e.g. they precede stops) are more confusable than fricatives (e.g. /f/, /v/, /s/, /z/) before oral stops, and this is reflected in the common neutralization of place contrasts[8] for nasals and

[8] A simple example of environment-specific *neutralization* can be found in many varieties of American English: in these varieties, the contrast between [t] and [d] is *neutralized* between stressed and unstressed vowels, meaning words like *medal* and *metal* or *ladder* and *latter* are not pronounced distinctly, even though speakers of these varieties productively distinguish [t] from [d] in other environments.

stops before stops; comparatively, neutralization of place contrasts among fricatives before stops is uncommon (Hura, Lindblom, & Diehl, 1992; Kohler, 1990; J. J. Ohala, 1990a). Patterns like this in phonology have been dubbed *perceptibility effects* (Steriade, 2001b). Two summary statements about such patterns are below:

(2.2.8) Languages tend not to make use of phonological contrasts in phonotactic contexts where they are difficult to perceive.

(2.2.9) Alternations[9] that result in *less* perceptible changes are more common than those that cause *more* perceptible changes.

To clarify (2.2.8): Suppose some language's lexicon uses /n/ and /m/ *contrastively*, i.e. there exist wordforms in that language's lexicon that are distinguished by the use of /n/ and /m/ – say two stems with a common prefix and that end in /an/ and /am/. Empirically, such a language is *less likely* than would otherwise be expected to make use of the [n-m] contrast in a context like [vowel__voiceless obstruent] where that contrast is difficult to accurately perceive.

Continuing the example, suppose there exists a *suffix* whose typical realization is /pa/ (i.e. it starts with a voiceless obstruent) and that can attach to both of the stems ending in /an/ and /am/. Naively, we might expect that both /…anpa/ and /…ampa/ are attested forms; in fact, more often than we might otherwise expect, /…an/+/pa/ does **not** manifest as /…anpa/.

If our example language doesn't have /…anpa/,[10] what does it have instead? Is *any* kind of 'repair' as likely as any other? Empirically (Steriade, 2001b), per (2.2.9), the answer appears to be *No.* Instead, we expect (ceteris paribus) that rather than an /n/, /…an/+/pa/ will make use of a segment type in the language's inventory that is perceptually difficult to

[9]The kind of categorical and predictable variation in the phonetic realization of the English plural ([s] vs. [z], etc.) and /p/ described in §2.2.1 are both examples of alternations.

[10]I.e. our example language doesn't combine /…an/+/pa/ as /…anpa/.

distinguish from /n/ – e.g. an /m/. I.e. **instead of** /an/+/pa/→/anpa/, the language is likely to have /a**m**pa/.[11]

This same pattern also plays out in language change (J. J. Ohala, 1990a), as exemplified by these changes in the history of French:

(2.2.10) a. Late Latin *primu tempus* ('first time, first season') > Old French *printans* ('spring')

b. Latin *amita* 'paternal aunt' > Old French *ante*

Other examples of perceptibility effects include

(2.2.11) a. neutralization of consonant voicing contrasts before other consonants (Hansson, 2008)

b. nasalization of vowels next to laryngeal sounds like [h] (J. J. Ohala, 1975)

c. stop insertion between nasals and following consonants (J. J. Ohala, 1974)

d. place of articulation shifts in fricatives (Hansson, 2008)

e. velar stops (e.g./g/, /k/) becoming labial (e.g./b/, /p/) (J. J. Ohala, 1993) or alveolar (e.g./d/, /t/; Chang, Plauche, and Ohala 2001)

among many others. See Blevins (2008, §2.2.1) for a centralized list of such effects.

In the following subsections, I review existing explanations of perceptibility effects in more detail, dividing them principally into

- *diachronic-phonetic* accounts that only articulate a role for listener-speakers at small scales and processes of conventionalization in large ones
- *synchronic-phonological* accounts that claim an explanatory role for direct representation of phonetic facts in the grammatical knowledge of speakers.

[11]The reason why we expect specifically assimilation of the /n/ (articulated at the alveolar ridge behind the front teeth) forward to the place of the /p/ (involving the lips) rather than the reverse – why /…ampa/ rather than /…anta/ – is the asymmetry in confusability of nasals in consonant clusters explained in the paragraph above Ex. 2.2.5.

2.2.3 Diachronic-phonetic accounts

While the focus of this chapter is a critical evaluation of synchronic-phonological explanations of perceptibility effects, the relative paucity of explicit rhetorical clash between synchronic-phonological accounts of perceptibility effects and diachronic-phonetic ones means that the redundancy, systematic misprediction of variation, and *a priori* implausibility of arguments for synchronic-phonological explanations cannot be appreciated without a review of diachronic-phonetic accounts of perceptibility effects.

Contemporary diachronic-phonetic accounts can be divided into two main groups. I discuss them in approximate chronological order. The key point of this subsection is that the two main groups of diachronic-phonetic accounts disagree about whether perceptibility effects are or ought to be explainable with reference to communicative pressures.

2.2.3.1 Listener-error accounts

The uncontroversial claim of what I will call 'listener-error' accounts (whose contemporary exploration more or less begins with the work of Ohala – see e.g. J. J. Ohala 1981, 1993) is that phonetic processes (e.g. coarticulation[12]) introduce a 'pool of [phonetic] variation' in how a stable, categorical object like a phoneme is realized. For example, while English doesn't have contrastively nasal vowels, pre-nasal vowels in American English are categorically nasalized by some speakers and variably ('phonetically') nasalized by the rest (Beddor, 2009).[13] The variation around each such phonological category is, in general, capable of much ambiguity and overlapping with variation resulting from other phonological categories: the speech signal underdetermines a speaker's intended representation. The

[12]While symbolic transcriptions of speech indicate a linear, segmented sequence of discrete speech sounds — e.g. [blɪk] — each speech sound is produced by continuous, dynamic, and parallel motor gestures of multiple articulators that temporally overlap. 'Coarticulation' refers to the partially overlapping, contextually sensitive, and interacting articulation of two or more temporally adjacent speech sounds.

[13]To convince yourself of this, pinch your nose closed while saying *bet* or *back* and compare this experience with doing so while saying *bent* or *bank*.

main claims of listener-error accounts (e.g. Blevins, 2004; Bybee, 2001; Garrett, 2015; J. J. Ohala, 1981, 1993) continue from here by observing that listeners may make *mistakes* ('innocent misapprehensions') about the cause of the variation they perceive (attributing something different to the speech sound than what the speaker intended) and then reproduce novel variants incorporating such 'mistakes'. As a result of this, a lexicon with contrasts that are difficult to accurately perceive or learn will, unsurprisingly, tend to not be accurately perceived or learned. This is the listener-error answer to the Mechanism Question.

In contrast to phonological accounts, these facts are taken by listener-error accounts to be statements about *speech perception* and its consequences for *language change*, **not** facts that are directly part of the productive knowledge humans have about patterns governing speech sound sequences (i.e. phonology) in the same way, that for example, an English speaker encountering a new noun like /blɪk/ (*blick*) or /wʌg/ (*wug*) expects that the plural is [blɪks] and [wʌgz] rather than *[blɪkz] or *[wʌgs]. As far as listener-error accounts are concerned, learners simply learn the arbitrary patterns that history and their environment give them, subject to the asymmetric filter of perception: nothing about the representational content or inferred generalizations about sound patterns associated with phonological perceptibility effects makes direct reference to facts about perceptibility.

A representative example of a common sound change and a listener-error account of it is the emergence of phonemically nasal vowels. As a result of some mix of coarticulation and confusability, pre-nasal vowels are often (phonetically) nasalized, as mentioned above: underlying /VN/ sequences are often produced as [$\widetilde{\text{V}}$N]. Together with e.g. reduction of the nasal segment, a listener could easily hear the acoustics associated with [$\widetilde{\text{V}}$N] and infer that /$\widetilde{\text{V}}$/ was the representation the speaker intended – i.e. that nasalization is phonemic and part of lexical representation rather than an incidental (if systematic) coarticulatory artifact (J. J. Ohala, 1992, 1993). Such changes have given rise to phonemically nasal vowels in a number of Romance language varieties (e.g. French, Beddor 2009; J. J. Ohala 1990a;

northern Italian varieties and Romanian, Sampson 1999) and Early Proto-Slavic (Padgett, 1997); in fact, most languages with phonemically nasal vowels are currently believed to have acquired their nasal vowels through such a process (Beddor, 2009).

As suggested by Figure 2.1, a more precise name for the class of explanations for perceptibility effects described here might be listener-error-*only* accounts: as detailed in §2.2.3.2, these accounts deny any significant explanatory role for communicative pressures.

2.2.3.2 Speaker-choice accounts

The second main diachronic-phonetic account of perceptibility effects (e.g. J. Kirby, 2013; Lindblom, 1990; Lindblom et al., 1995) agrees with the listener-error account, but elaborates on the listener-error answer to the Mechanism Question: essentially, it claims that the variation listeners are exposed to is shaped by the choices of speakers, who (according to the hypothesis) are seeking to achieve a comfortable balance of ensuring that useful acoustic cues reach the listener while not expending unnecessary articulatory effort.

The contemporary motivation for these accounts of perceptibility effects can be associated with the work of Lindblom, work principally aimed not per se at explaining language change but rather offering a hypothesis for the character of online variation in speech production. That is, speech is highly variable, and decades' worth of searching for relatively simple invariant acoustic (and/or articulatory) cues to the identity of speech sounds (used by humans or usable by machines) in the mid to late 20th century came up empty handed (Carbonell and Lotto see e.g. 2014 for an overview). Instead, there are, in general, many possible phonetic cues to any given contrast, varying in kind and quality across contexts. Lisker (1986), for instance, offers more than a *dozen* different cues just to the voicing distinction among English oral stops (/p/ vs. /b/, /t/ vs. /d/, /k/ vs. /g/, etc.).

Lindblom (1990)'s *H & H theory* hypothesized that a large proportion of variation in the speech signal is a result of speakers *hypo-articulating* those acoustic cues they think

they can get away without (i.e. those that aren't important to the listener inferring what the speaker wants to be reasonably confident will be transmitted) and *hyper-articulating* those cues that are necessary or important, based on some model of the situation, the speaker's goals, and a model of the listener.[14] This research program has been fairly successful in formalizing Lindblom's ideas and experimentally testing them (see e.g. Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Buz, Tanenhaus, & Jaeger, 2016a; Hall, Hume, Jaeger, & Wedel, 2018; Seyfarth, 2014; Van Son & Pols, 2003, among others). Cohen Priva (2012) is notable for connecting this to conversational phonological/phonetic processes and explaining language-specific variation not accounted for by the synchronic-phonological accounts of perceptibility effects discussed later.

Returning to language change, speaker-choice accounts are, in fact, compatible with listener errors as ***a*** source of sound change: as briefly mentioned above, where speaker-choice accounts primarily differ is that *they explore the role of speaker production choices in shaping the variation that listeners are exposed to in the first place*, and in particular how the differential hyperarticulation of some acoustic cues and the hypoarticulation of others may explain the direction of listener reanalysis and therefore ultimately phonological change. This means that speaker-choice centered accounts of perceptibility effects

- draw on work with an independent motivation – seeking to explain production variation (not centrally motivated by explaining language change).
- attempt to answer questions about perception and change unexplained by listener-only accounts.

[14] See Degen (2013, p. 43-45) and Piantadosi, Tily, and Gibson (2012, §2) for discussions at other levels of the linguistic hierarchy, theoretical arguments, and reviews of corpus and experimental evidence for the idea that context is informative, listening/comprehension is cheap, production is expensive, and that therefore understanding the *problems* that production choices solve (and how well different choices solve them) is crucial for explaining online variation in what choices speakers actually make and what choices grammars offer to speakers in the first place.

What makes this class of diachronic accounts continue to remain distinct from Ohala-esque listener-error(-only) accounts is the *communicatively-adaptive* explanation they offer that many researchers regard as some mix of a priori implausible[15] and/or abhorrent:[16] 'Sound change is not teleological; it does not serve to optimize articulation, perception, or the way language is processed in the speaker's brain. It is just an inadvertent error on the part of listeners.' (J. J. Ohala, 1997)

2.2.3.3 Evaluation and relation to the present work

With respect to the Mechanism Question, speaker-choice accounts agree with listener-error accounts that systematic biases in misperception are part of the explanation for perceptibility effects, but extend this by claiming that the pool of variation from which mistakes are drawn is shaped by speakers actively trying to enhance important phonetic cues and attenuate cues they can get away with attenuating. Neither class of diachronic-phonetic accounts accords direct representation of phonetic facts in phonology an explanatory role for perceptibility effects.

Across the board, diachronic-phonetic accounts have at least three strengths. Unsurprisingly, phonetic accounts excel at describing specific, acoustically and/or articulatorily plausible conditions under which misperceptions could occur that could give rise to well-attested sound changes and not to rare or unattested ones.

Second, phonetic-diachronic accounts emphasize the accidental and contingent

[15] As Carbonell and Lotto (2014) elaborates for the case of speech, a running theme throughout the first several decades of the cognitive sciences everywhere but the study of low-level perceptual processes (Green & Swets, 1966) is the assumption that human learning and inference capacities *must* be suboptimal because they didn't match (in hindsight naive) intuitions about what normatively intelligent behavior should be and because early general formulations of intelligence (Newell & Simon, 1961; Robinson, 1965) were quickly understood to be impractical for real world problems (see e.g. Newell, 1981, p. 3).

[16] In both the context of explaining online variation (Gahl, Yao, and Johnson see e.g. 2012) and specifically in the context of language change (Bybee see 2001 or discussion of Bybee's work in Hansson 2008), there are non-adaptive but production-oriented accounts of change that center around explanations for the relationship between reduction and frequency effects, but see review and critical discussion in Hall et al. (2018, §2) and Jaeger and Buz (2017).

nature of listener reanalysis events; i.e. unlike a grammatical (and especially an innate grammatical bias) account, phonetic-diachronic accounts don't need any special qualification or explanation to be statistical *tendencies*. Empirically, this means that they can point to the existence and learnability of so-called 'unnatural' sound changes as evidence against strong nativist formulations of synchronic-phonological accounts (see e.g. Blevins, 2008) that overfit well-known data.

Finally — and crucially for this chapter — phonetic accounts are highly compatible with contextual variation within a language: acoustic cues and behavior of both listeners and speakers are known to involve a large amount of variation even within a given language (see e.g. Keating, 1985). Further, (as elaborated in the next full section) the conditions under which a particular segment in a particular environment is easy (or hard) to correctly identify should be expected to vary as a function of context.

2.2.4 Synchronic-phonological accounts

Similar to diachronic-phonetic accounts, synchronic-phonological accounts also divide into two main groups, which I again introduce in chronological order. While the first group is not the focus of this chapter, I cover them nevertheless to underscore how weak of a rhetorical position synchronic-phonological accounts are in with respect to explaining perceptibility effects — a conclusion not derivable from an isolated review of the subset of the synchronic-phonological literature most relevant to this chapter.

The five key points of this subsection, in descending order of importance, are:

- There is a key implicit methodological assumption concerning the specific kind of phonetic facts about perceptibility these accounts incorporate into phonological representations that is common across all synchronic-phonological accounts and that is central to my criticism of them: the measure of relative confusability of speech sounds used only takes the acoustics of a speech sound and its local phonotactic

environment into account. The incompleteness of this assumption is elaborated in following sections, where the psycholinguistics of comprehension and recognition are reviewed and mathematically explicated.

- No synchronic-phonological accounts dispute that diachronic-phonetic processes play at least some role in explaining perceptibility effects.
- The main division among synchronic-phonological accounts concerns their stance on the role of communicative pressures: early (but not later) work explicitly motivated grammatical devices for explaining perceptibility effects in terms of communicative pressures on speakers.
- Until the last two decades, most work synchronic-phonological work was silent on the causal or explanatory role of phonological models of perceptibility effects.
- Recent work posits an *innate* and specifically phonological *bias* for more perceptible sound sequences and less perceptibly salient alternations.

The reader satisfied with this may skim or skip this subsection; the next subsection contains a summary of both diachronic-phonetic and synchronic-phonological accounts and key points about them in an intermediate level of detail.

2.2.4.1 Communicative accounts

2.2.4.1.1 Key claims

The first set of synchronic-phonological accounts (e.g. Flemming, 2001a; Jun, 1995; Padgett, 1997) take Lindblom (1990) as their logical starting point: 'If phonological systems were seen as adaptations to universal performance constraints on speaking, listening, and learning to speak, what would they be like?' That is, the key features of their hypothesis are

- that perceptibility effects are explained by positing that phonological systems are

communicatively adaptive.

- ...and that this should be *directly represented*[17] in *synchronic phonological grammars* using Optimality Theoretic constraints.

Crucially, as will become clearer in the next sections, the constraints that these theories posit reference the relative and *contextual* confusability of different segments, where 'context', to date, has been an entirely *local* affair: the local context of a segment token refers at most to the immediately adjacent segment tokens – not e.g. the entire wordform. To be concrete: if $\rtimes x_0 x_1 x_2 x_3 \ldots \ltimes$[18] is a segment sequence ('wordform') actually produced by a speaker, an *entirely* out of context measure of confusability would be $p(\hat{x}_i | X_i = x_i)$ (or some function of this), the listener's beliefs about the speaker's intended ith segment given that the actual ith segment produced was x_i. A *local* measure of the confusability of x_i would be $p(\hat{x}_i | x_{i-k} \ldots x_i \ldots x_{i+k})$ for $k \in \mathbb{Z}^+$ close to 1. The fact that the notion of context considered to date is local in this sense and that it is psychologically inaccurate are critical for the main argument of this chapter — that synchronic-phonological approaches are implausible accounts of perceptibility effects.

In sum then, communicative synchronic-phonological accounts don't dispute diachronic-phonetic answers to the Mechanism Question, and they are explicitly motivated by much of what motivates the work of speaker-choice accounts. While they develop synchronic-phonological theories where facts about phonetics are directly incorporated and represented in what is taken to be a model of phonological competence (recall §2.2.1), they rarely write on why or whether this adds to explaining why perceptibility effects exist.

[17]That is, the intensional definition of constraints centrally responsible for perceptibility effects directly refer to facts about differential perceptibility of different segment sequences/segment-context pairs.

[18]I will use the $\rtimes$ and $\ltimes$ symbols to explicitly indicate the left and right edges of a word, respectively.

2.2.4.1.2 Formal proposals

In this section I offer an example of a formal analysis from communicatively-oriented synchronic-phonological accounts of perceptibility effects, preceded by a brief introduction to Optimality Theory. The purpose of this is to illustrate the nature of constraints and how synchronic-phonological accounts of perceptibility represent facts about relative and contextual confusability using them.

Communicative synchronic-phonological accounts began shortly after the advent of Optimality Theory (OT) (Prince & Smolensky, 1993). This is not an accident (Zhang, 2001, §1.4.1): OT's formulation of grammar as a set of ranked and violable constraints that well-formed strings must optimally satisfy lends itself to viewing phonology as a problem solving system and greatly facilitates describing common 'problems' (e.g. having perceptible words) that different languages face in a way that highlights what is shared across otherwise very different languages. That is, relative to earlier formalisms for describing phonological patterns (viz. the rewrite rule formalism of Chomsky and Halle 1968), OT constraints permit a straightforward means of describing patterns in strings that have an external motivation: well-formedness constraints can describe what properties make one wordform 'worse' than another, and the observed strings of a language are those that violate the most important constraints the least – i.e. the strings that represent optimal 'solutions'.

To briefly illustrate, consider loanword adaptation and non-native language learning: Spanish words are constrained such that they feature no word-initial consonant clusters[19] beginning with /s/; comparatively, English has no such restriction and in fact has many words that begin with such clusters. Instead of producing the English word /strɛs/ (*stress*) as [strɛs], then, a common phonological pattern among Spanish speakers who are not native English speakers is producing a very similar sequence of sounds – [estrɛs] – that satisfies

[19] Sequences of consecutive consonants uninterrupted by vowels.

this phonological restriction of Spanish (among others). Typical constraints in Optimality Theoretic analyses take one of two forms: 'Markedness' constraints, which penalize a specific set of output ('surface') strings if they have certain properties, and 'Faithfulness' constraints, which penalize all mappings from an input ('underlying') form to an output form that alter specific features of the input form. Markedness constraints can be usefully thought of as forces of change and Faithfulness constraints as forces of preservation. When two constraints conflict, the relative *ranking* of constraints determines which one prevails. In our example, Spanish could be analyzed as having a Markedness constraint against words with particular kinds of word-initial consonant clusters that outranks a Faithfulness constraint to e.g. not add segments that aren't in the original input. Typological diversity is explained by different languages having different rankings of the same (or more or less the same) constraints.

Work in what became known as 'phonetically-based phonology' ('PBP'; Hayes and Steriade 2004) represented hypothesized pressures for articulatory ease and perceptual contrast as constraints that

- penalize articulatorily costly sound segment sequences.
- penalize segments in phonotactic environments where they are difficult to accurately perceive.[20]

where relative penalization of different segment sequences (the relative ranking of constraints) follows relative articulatory cost or perceptual difficulty, respectively.

As an example, consider Jun (2004)'s analysis of place assimilation in consonant clusters. Jun begins with an example from the Niger-Congo language Diola-Fogny:

(2.2.12) /ni+gam+gam/ → [nigaŋgam] 'I judge'

[20]Alternatively: preferentially maintain more perceptible segment sequences/wordforms that are harder to correctly identify or discriminate.

That is, the three morphemes whose underlying representation is taken to be /ni/, /gam/, and /gam/, when combined to produce a single word are realized as [nigaŋgam]. Here, the place of articulation of /m/ (the lips) assimilates to the place of a following /g/ (the velum, or soft palate), but otherwise retains all the features of an /m/. As Jun notes, this is not an analytically difficult process to compactly describe, but nothing about the form or content of a traditional phonological rewrite rule or e.g. positing a constraint against consonant clusters with heterogenous places of articulation *explains*

- why such processes occur at all
- why such patterns are as common as they are
- why place assimilation is predominantly regressive[21]
- why nasals and coronals[22] seem to frequently be the sounds that are altered in an assimilation process
- or why non-coronals tend to be sounds that adjacent segments alter to assimilate towards.

Jun (2004)'s analysis is that these patterns can be explained with reference to differences in perceptibility of place cues among different segment types and how those cues are affected by different kinds of phonotactic environments.

Formally, Jun makes use of two types of phonetically-grounded constraints. The first, WEAKENING is an articulatory markedness constraint. As used by Jun, violations of it are assigned in proportion to the number of articulatory gestures added and the degree of articulatory closure of the vocal tract involved in them; it 'causes' weakening of the articulatory gesture, shortening of the sound, or even deletion altogether. The PRESERVE family

[21]Example 2.2.12 involves *regressive* assimilation; if instead, the underlying /m/ remained the same but the following /g/ changed only by assimilating to the place of the /m/ (i.e. as a [b], resulting in *[nigambam]), it would be *progressive* assimilation.

[22]Basically sounds produced near the front of the mouth, from the hard palate forward.

of faithfulness constraints penalize output candidates that remove or decrease perceptual cues to features of the input. The tableau below illustrates when assimilation rather than deletion occurs – when a constraint on preserving cues to manner outranks WEAK and a constraint on preserving cues to place.

Input = /mg/	PRES(MANNER)	WEAK	PRES(PLACE)
☞ a. ŋg (assimilation)		*	*
b. mg (no change)		**!	
c. g (deletion)	*!	*	*

Jun then proposes a universal ranking over the whole parameterizable family of PRESERVE constraints:

(2.2.13) PRES(X(Y)): Preserve perceptual cues for X (place or manner of articulation) of Y (a segmental class).
Universal ranking: PRES(M(N)) ≫ PRES(M(R)),
where N's perceptual cues for M are stronger than R's cues for M.

Jun then suggests rankings on particular PRESERVE constraints based on experimental data about the perceptibility of different sound classes, different cues within each class, and strikingly parallel implicational universals in the typology of place assimilation. For example, the 'target place' ranking,

(2.2.14) PRES(pl(dor˺)) ≫ PRES(pl(lab˺)) ≫ PRES(pl(cor˺))

is grounded in laboratory evidence that the place cues of unreleased dorsals[23] (e.g. those occurring cluster-initially) are more perceptible than those of unreleased labials[24], which in turn, are more perceptible than those of unreleased coronals. Typologically, there is a

[23]Sounds whose place of articulation is near the back of the oral cavity.
[24]Sounds produced with the lips.

corresponding implicational universal: dorsals are uncommon targets of place assimilation, and when they are attested, labials and coronals also are; labials are somewhat more common, and when they are attested, coronals also are. Accordingly, typological variation comes from different languages interleaving a WEAKENING constraint at different points in the universal ranking Jun proposes:

(2.2.15) a. **WEAKENING** ≫ PRES(pl(dor)) ≫ PRES(pl(lab)) ≫ PRES(pl(cor))
→ Coronals and noncoronals are all targets.

b. PRES(pl(dor)) ≫ **WEAKENING** ≫ PRES(pl(lab)) ≫ PRES(pl(cor))
→ Labials and coronals are targets but velars are not.

c. PRES(pl(dor)) ≫ PRES(pl(lab)) ≫ **WEAKENING** ≫ PRES(pl(cor))
→ Only coronals are targets.

Note that in this case, perceptibility constraints make no direct reference at all to phonotactic context.

Besides consonant cluster assimilation, some examples of other phenomena addressed by communicative synchronic-phonological accounts include

- vowel inventories (Flemming, 2001a)
- neutralization of vowel height contrast following nasalization (Padgett, 1997)
- contrastive palatalization emerging out of the loss of following high vowels ('yer deletion') (Padgett, 2003)
- rhotic contrast neutralization (Bradley, 2001; Padgett, 2009)
- chain shifts (Łubowicz, 2003)
- contrast preservation effects in morphological paradigms (Łubowicz, 2007).

In sum, as Jun's analysis of place assimilation in consonant clusters illustrates, the attraction of PBP (particularly formalized in OT) to phonology is that it offers an external source of

evidence for constructing and evaluating explanations of variation in phonological typology that is less stipulative than alternatives.

Finally, it is worth noting that the formal mechanisms of much of this work (e.g. Flemming, 2001a; Łubowicz, 2003; Padgett, 1997, 2003) go much further than Jun, arguing that rather than just capturing a preference for words whose *sounds* are *more perceptible*, ultimately what matters is that *whole words* be *distinguishable* – i.e. that contrast should be analyzed at the level of whole words. Accordingly, they argue that where traditional OT models the phonological patterns of a language via a calculation process that can separably and in parallel operate to produce each individual wordform, these authors argue instead for a theoretical architecture with *global, systemic optimization* of sound inventories and lexicons.

2.2.4.1.3 Relationship to diachronic-phonetic accounts

Communicatively-oriented synchronic-phonological work cites the phonetic literature offering mechanisms for perceptibility effects. For example,

- Jun (1995, pp. 2, 29) references the Production Hypothesis: 'More articulatory effort is likely to be invested in the production of sounds with powerful acoustic cues than those with weak cues.' Jun also cites Lindblom's H&H theory (Jun, pp. 27, 156).
- Flemming (2001a, pp. 15-16) similarly references Lindblom's theories.

What then is the relationship of communicatively-oriented synchronic-phonological accounts to the communicatively-oriented analogue work in phonetics it cites? There are two salient options:[25]

(2.2.16) The grammatical analyses synchronic-phonological work advocates are convenient formalisms aiding linguists in compactly and insightfully describing patterns

[25]Compare with Scholz, Pelletier, and Pullum (2011)'s *Externalists* and *Essentialists*.

in speech sound sequences in terms of structure-external causes that shaped them.

or

(2.2.17) Per traditional generative theory, formal theories of a language's grammar are

a. A description in some sense of the productive knowledge and representations in speaker's heads.

b. Distinct or distinguishable from the the implementational mechanisms of learning, producing, and comprehending language.

(2.2.16) poses no conflict with diachronic-phonetic accounts of perceptibility effects. (2.2.17), however, is a strong claim about the mental representations of productive phonological generalizations associated with perceptibility effects that is difficult to reconcile with diachronic-phonetic accounts (communicatively-oriented or not), particularly given (2.2.17b). That is, given that communicatively-oriented synchronic-phonological accounts *don't* dispute the explanations that diachronic-phonetic accounts offer for how perceptibility effects arise, if authors like Jun or Flemming *were* to argue for (2.2.17), it is not clear what phonetically-motivated constraints explain that isn't *already* explained by phonetic accounts and that a phonological theory with substantially simpler constraint types couldn't capture. Why – and how – should speakers/learners model the relative perceptibility of different wordforms (including unattested ones) in the context of *phonological* (rather than phonetic) representations?

To my knowledge, most authors in the communicatively-oriented synchronic-phonological literature (e.g. Jun and Padgett) take no stance on this question. Flemming's only explicit position is agnostic (p.c., also Flemming 2001b, fn. 9), but suggests looking to the psycholinguistic literature for future evidence. Boersma (1998) argues for (2.2.17), inclusive of (2.2.17a) but explicitly rejecting (2.2.17b). As described below, the perceived lack of a theory of phonetically-grounded constraints with compelling psychological reality

and explanatory value (and plausibly apprehension about the analytic unwieldiness of more systemic contrast optimization proposals as in e.g. Padgett 2003) continued into the early 2000s and left a void in phonetically-based phonology filled a new line of research in the other main class of synchronic-phonological explanations for perceptibility effects.

2.2.4.2 Similarity-based accounts

2.2.4.2.1 Key claims

The key works in the second class of synchronic-phonological accounts are Steriade (2001a) and Steriade (2001b).[26] Its key claims are that

- Humans have *grammatical knowledge* – a so-called *perceptibility map* (or 'P-map') of …
- … the relative contextual *perceptual similarity* of different segment types.

and that

- This knowledge is directly translated into phonological generalizations (OT constraints).
- This *grammatical* knowledge that individuals possess is taken to explain why phonological processes (e.g. alternations) that involve a less perceptually salient change (relative to an underlying form) are more likely than otherwise plausible alternatives that involve more salient changes.

Note that 'similarity' at the theoretical level was (and continues to be) left vague,[27] but that, in practice, (local) measures of segmental confusability have been deemed a convenient operationalization (Hayes & White, 2013; Steriade, 2001a, 2001b) of this similarity for

[26] A revised version of this second manuscript was published as Steriade (2008).
[27] See Gallagher (2012).

analysts.[28]

In sum then, similarity-based accounts don't dispute the diachronic-phonetic answer to the Mechanism Question and the enthusiasm present in earlier work in the synchronic-phonological literature for communicative adaptation is absent.

2.2.4.2.2 Formal proposal

The substance of Steriade's proposal is that a speaker's knowledge of relative perceptual similarity maps directly to how faithfulness constraints should be ranked: more distinctive contrasts are preferentially preserved over less distinctive contrasts. For example, listeners find it easier to distinguish [ba] from [pa] than [abta] from [apta] (Malécot 1958, Wright 2004, §2.1.3):

(2.2.18) Δ(b,p) >> in context [⋊ __vowel] than [vowel__voiceless obstruent].

Or, compactly,

(2.2.19) [b-p]/⋊_V >> [b-p]/V_C.

The corresponding ranking of faithfulness constraints[29] is

(2.2.20) IDENT(voice)/⋊_V >> IDENT(voice)/V_C.

i.e., that changes to voicing are penalized more harshly in the environment ⋊_V than in the environment V_C.

2.2.4.2.3 Substantively-biased phonology

Early work by e.g. Hayes (1999) (a precursor to Steriade's P-map), Steriade (2001a), and Steriade (2001b, 2008) explicitly emphasize the role of *experience* in develop-

[28] Alternative or additional sources of evidence for acoustic similarity in this body of literature include attested usage of literary devices like puns, rhymes, and alliteration where perceptual similarity is part of the literary device.

[29] Note that e.g. if [g-k]/⋊_V >> [g-k]/V_C were also true, it would translate to the same constraint.

ing knowledge of relative perceptual similarity and related phonetic knowledge. Within a few years, however, so-called *substantively-biased phonology* ('SBP'; centrally Wilson 2006, but see also e.g. Hayes and White 2013; Moreton 2008; Moreton and Pater 2012a, 2012b; White 2014, 2017; Zuraw 2007), hypothesized that learners may have an *innate bias* that predisposes them to learning 'phonetically natural' phonological patterns, but that this bias is 'soft' – overrideable by experience. Though some work in this subgenre of similarity based accounts has introduced innovations in constraint representation (Zuraw, 2007, 2013), the main distinguishing feature of substantively-biased phonology from Steriade's original proposal is its claim about the origin of the P-map.

This move – exploring the idea that some portion of the P-map is *innate* – was motivated in part by the perceived failure of earlier work (both by Steriade and earlier synchronic-phonological accounts) to make a strong case for the *necessity* of a synchronic-phonological analysis of perceptibility effects.[30]

2.2.4.3 Evaluation and relation to the present work

The main apparent weaknesses of synchronic-phonological center around the unclear explanatory value added by inserting direct reference to phonetic facts into grammatical knowledge. Most synchronic-phonological work on perceptibility effects is silent on this question, but if any are taken to have meaningful correspondence to psychological reality, then the Mechanism Question also becomes problematic: when is constraint optimization supposed to happen, particularly for accounts arguing for systemic optimization of the entire lexicon (or significant fractions of it) simultaneously?

Synchronic-phonological accounts arguing for an innate representational bias for functional patterns ('substantively-biased phonology', §2.2.4.2.3) must also defend

[30]See §1.2 of Zhang and Lai (2006) or Zhang and Lai (2010) for a clear discussion of the issues and critical summary of evidence; with respect to §1.3's analysis of the utility and interpretation of learning experiments – like that of Wilson (2006) and later work – see §2.2.4.3.

against both critiques from nativists (M. Hale & Reiss, 2000; Reiss, 2017) that phonology is autonomous and strictly separated from phonetics, as well as from empiricists (see e.g. Baronchelli, Chater, Pastor-Satorras, & Christiansen, 2012; Chater, Reali, & Christiansen, 2009; Elman et al., 1996) who argue that innate specification of fine-grained representations about a domain as evolutionarily recent and complex as language is implausible, especially given increasing evidence for the powerful learning capabilities of children. Worse, the primary method proposed for finding evidence of a substantive inductive bias has had at best mixed results (Moreton & Pater, 2012a, 2012b).

2.2.5 Summary of previous work

The two main branches of explanations for perceptibility effects are diachronic-phonetic and synchronic-phonological.

Diachronic-phonetic accounts of perceptibility effects answer the Mechanism Question by arguing that what explains perceptibility effects in phonology are processes of speech perception, lexicon/phonological learning, and propagation of phonological conventions through populations. Diachronic-phonetic accounts can also be further divided into (at least) two groups, divided principally by their stance on the role of communicative pressures in explaining perceptibility effects.

The first group – 'listener-error' accounts – articulates an online mechanism by which listeners reanalyze ('innocently misapprehend') typically ambiguous acoustic data as being caused by linguistic structures potentially different from those actually intended by speakers. Systematic differences (subject to some phonetically describable language-specific variation) in what contrasts are easy to perceive (in what phonotactic contexts) and what segment types are easy to correctly identify and learn (and again, in what phonotactic contexts) are argued to explain the cross-linguistic prevalence of perceptibility effects in the phonologies and historical trajectories of natural languages.

The second body of work ('speaker-choice' accounts) extends the first, but goes further by hypothesizing that the ambiguous and highly variable acoustics that listeners must infer and learn linguistic representations from is not randomly variable: its distribution can be explained with reference to adaptive online choices by speakers that reflect their specific communicative goals, a model of the situation, and a model of the listener. That is, this work argues that speakers adaptively choose to enhance phonetic cues that are useful for helping the listener correctly understand them and ameliorate phonetic cues to the extent that doing so makes articulation easier without significantly impacting expected communicative success.

Diachronic-phonetic accounts have gone to great lengths to elucidate the conditions under which particular sounds and environments are likely to be misperceived and how, and to relate these facts about phonetics to well-attested historical patterns of change. Nevertheless, there is plenty of room for elaboration in answering the Mechanism Question: considerably less well studied are theories and explanations of *phonologization* of misperceptions within individuals or how these spread within populations. That is, it is an open problem how one or more misperceptions of what a speaker intended result in a change to an individual's phonological grammar, and what the conditions, mechanisms, and timescales are by which such misperceptions spread through the community of a language variety. Similarly, perceptibility is not the only logically possible explanation for phonological change in general, and it is not clear that perceptibility effects in phonology have effects distinguishable from other causes (e.g. sociolinguistic factors), nor is it clear what the relative explanatory strength of perceptibility is relative to other factors.

Synchronic-phonological accounts of perceptibility effects argue that phonological grammars should contain Optimality Theoretic constraints that directly reference the relative discriminability of different speech sounds in different contexts. This literature contains two groups of proposals, distinguished chiefly by their stance on the explanatory role of

communicative pressures.

The first group argues that incorporating perceptibility constraints into grammars reflects the effect of communicative pressures on the organization of sound patterns, but remains silent or agnostic about the psychological reality of these proposals as statements about the *grammatical knowledge* of speakers.

The second group ('similarity-based accounts') argues that incorporating perceptibility constraints into grammars reflects knowledge that individuals have about possible, preferable, or more likely phonological patterns ('processes'): less perceptually salient 'changes' (relative to posited underlying forms) are preferable to plausible alternatives that involve more perceptually salient changes.

Crucially for this chapter, both literatures make use of behavioral results and quantitative measures of contextual confusability that use a highly local and strictly phonotactic notion of context (nearby segments). The second body of work, however, views confusability measures as a convenient operationalization of perceptual 'similarity' among a small set of tools for determining 'similarity', i.e. confusability measures permit much broader and more quantitative coverage of phonological inventories than alternative methods about what remains an otherwise nebulously defined notion of 'similarity'.

As explanations of perceptibility effects, the main strength of phonological analyses is that they permit relatively explicit (compared to phonetic-diachronic accounts) examination and comparison of perceptibility effects with other phonological generalizations. Specifically communicatively-oriented phonological work has also made it clear that to analyze contrast requires phonological architectures with simultaneous consideration of many sets of wordforms in the lexicon rather than separable computation of each wordform; this foreshadows my argument that accurately modeling perceptibility and communicative pressures requires even more substantial revision than imagined by this subset of communicatively-oriented phonological work.

Finally, synchronic accounts are lacking, however, in offering an explanation for how facts about online perception end up as perceptibility-related constraints with the rankings they have (answering the Mechanism Question), as well as what explanatory purpose is served by directly representing facts about perceptibility in grammars if there is a clear causal explanation for perceptibility effects in speech perception and patterns of language change. The notable exception to this is a contemporary subset of similarity-based accounts ('substantively-biased phonology') that posits the existence of an innate and specifically phonological bias for phonological alternations that make less perceptually salient changes.

2.3 Context and psycholinguistic processing

The core of my argument against extant phonological accounts of perceptibility effects is that such work has relied on a notion of segmental confusability that is incompatible with what is known about

- the *psychology of cue combination and inference*
- *variation* in the stable structure of non-local *communicative context* that speech sounds appear in.

Existing synchronic-phonological accounts use what I will interchangeably call *out-of-(global)-context* confusability, *out-of-context* confusability, or confusability that only references *local context*: i.e. they use measures of confusability where the *only* notion of context that goes into calculating the confusability of a segment token is the nearby segmental ('phonotactic') context. I will refer to and defend the normatively and psychologically motivated use of what I will interchangeably call *in-(global)-context* confusability, *in-context* confusability, or confusability that references *global* epistemic context: such measures not only refer to bottom-up perceptual information (including local co-articulatory effects), but

combine this with incrementally updated top-down expectations.

In this section, I review existing psycholinguistic work; in the following section, I present a computational model of a relatively simple task that permits examination of relatively simple cases of *inference* and *variation in communicative context* that are adjacent to phonetics and phonology: the psycholinguistics of *incremental spoken word recognition* and a *lexicon*, respectively.

2.3.1 Integration of top-down expectations and bottom-up perceptual cues

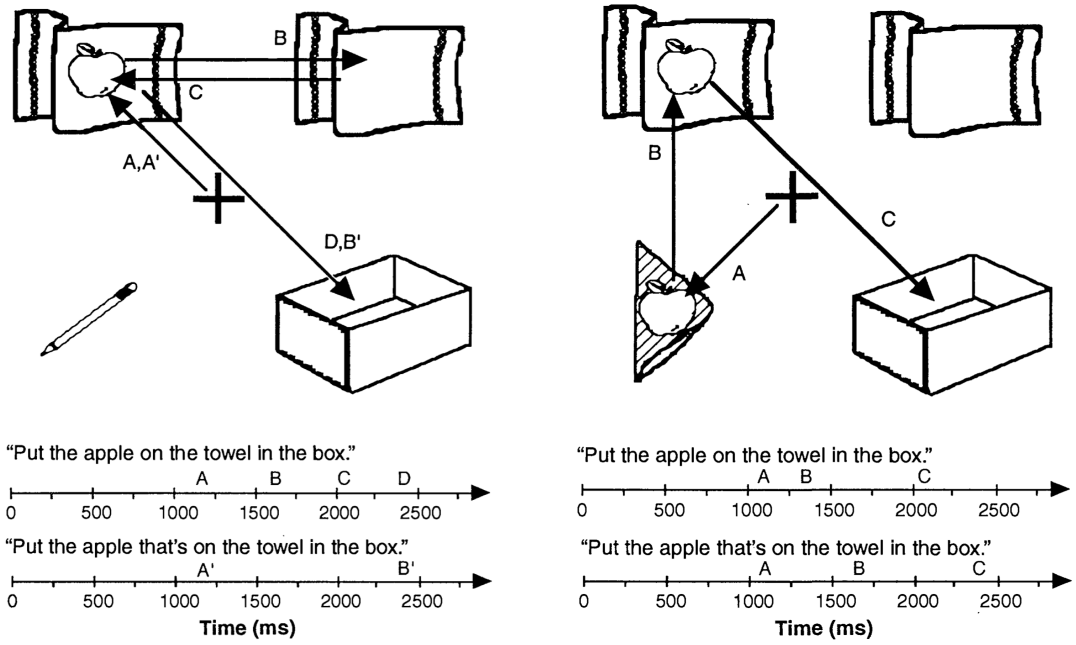
Recall from earlier that both classes of synchronic accounts have used speech sound confusability measures to define or operationalize perceptibility or 'perceptual similarity' constraints that at most take into account the effect of adjacent sound segments on perceptibility. Because of the distance between psycholinguistics and phonetics/phonology as research communities (at least historically) and the fact that confusion matrices have typically not manipulated phonotactic context for practical reasons, there is no reason to think this was a conscious choice.

Unfortunately, this notion of confusability is specifically known to be false in the case of psycholinguistics.[31] Relatively early work on language processing suggested that listeners understand utterances and words incrementally and that contextual information (e.g. previously read or heard words in the current sentence or previous sentences) robustly facilitates language processing at a variety of levels of the linguistic hierarchy – see Morton (1969), Marslen-Wilson (1973), or Marslen-Wilson and Tyler (1980) for early work at the word level, Marslen-Wilson (1975), Marslen-Wilson, Tyler, and Seidenberg (1976), or Tyler

[31] In fact, it is generally inaccurate across other domains of human cognition – see Knill and Richards (1996) for a collection of contemporary work, or e.g. Jacobs and Kruschke (2011) or Clark (2013) for more recent review and discussion of cue combination (among other topics) across research on perception, motor cognition, categorization, and learning.

and Marslen-Wilson (1977) at the sentence level, or the theoretically-oriented summary in Marslen-Wilson and Tyler (1987). For example, the ability of listeners in shadowing tasks to nearly simultaneously echo spoken passages as they listened, the robust ability of listeners in shadowing tasks to predictively correct experimenter-inserted errors (Marslen-Wilson, 1975) and the ability of listeners to accurately guess what word is being said in gating tasks (Grosjean, 1980; Wayland, Wingfield, & Goodglass, 1989) based on less than half of the total acoustic input provided early evidence that listeners (1) do not require all acoustic information about a word before they are able to accurately infer what word is being said and (2) that listeners not only make use of perceptual information within the current word, but also use at least some kinds of linguistic information from other parts of context in the course of word processing. Incremental cue integration, unsurprisingly, then, has long been a central part of models of spoken word recognition (e.g. Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Morton, 1969; Norris, 1994).

Perhaps the most dramatic example illustrating how manipulation of incrementally-updated top-down expectations of listeners can change the interpretation of units of form is given by Tanenhaus et al. (1995). In this study, participants looked at a visual scene on a computer screen while having their gaze tracked, listened to instructions about how to manipulate objects in the scene, and then carried out the instructions. On critical trials, these instructions either did or (in the control condition) did not contain a structural ambiguity that could change the action participants should do: *Put the apple on the towel in the box.* vs. *Put the apple that's on the towel in the box.* Previous work showed a brief, consistent preference by participants for one structural interpretation (comprehending the first prepositional phrase as a destination rather than a modifier) even when previous *linguistic* context supported an alternative interpretation. Fixation patterns in Fig. 2.2a show evidence of a listener preference for initially interpreting the first prepositional phrase as a destination rather than a modifier, though they eventually recover. This was interpreted as evidence for syntactic



(a) The unambiguous world condition.

(b) The ambiguous world condition.

Figure 2.2: The timecourse of fixations in Tanenhaus et al. (1995) across both manipulations. (From Figures 1-2 of Tanenhaus et al. (1995). Reprinted with permission from AAAS.)

modularity – insulation of early syntactic processing from other sources of information and a prior preference for structural simplicity in parsing, attributed to the computational difficulty of parsing.

Tanenhaus et al. hypothesized that contextual information may be delayed via working memory processes from being accessible during sentence comprehension; using immediately accessible *visual* context might demonstrate a clear effect of contextual knowledge on sentence processing and therefore that the comprehension process does not in general feature either the strong encapsulation or the difficulty in integrating heterogenous sources of information argued to be necessitated by a modular computational architecture for language processing. Accordingly, in addition to manipulating the structural ambiguity of the sentence on key trials, a second manipulation was whether visual context afforded

one referent (disambiguating world context) for the incrementally ambiguous portion of the sentence input vs. two referents (ambiguous world context). Crucially, fixation patterns indicate that in the disambiguating world context, listeners had no trouble interpreting the initial prepositional phrase as a modifier rather than as a destination. (Compare Fig. 2.2a with 2.2b.) The significance of Tanenhaus et al. (1995) is that it illustrated that even non-linguistic top-down cues about situation-specific world knowledge can be rapidly integrated and combined by listeners with bottom-up linguistic cues: the scope of 'context' that affects language processing is vast and extends far beyond even a presently-unfolding wordform.

Within the last decade or so, and reflecting a more general trend of research into modeling and explaining human cognition as an approximately rational set of computational processes for solving problems of inference, learning, and decision-making (Anderson, 1991; Chater & Oaksford, 1999; Griffiths, Kemp, & Tenenbaum, 2008),[32] recent work in language comprehension (e.g. Levy, 2008a; N. J. Smith & Levy, 2013) in general and (both aural and visual) word and speech processing in particular (see e.g. Clayards, 2008; Feldman & Griffiths, 2009; Feldman, Griffiths, & Morgan, 2009; J. Kirby, 2013; Kleinschmidt & Jaeger, 2015; McQueen & Huettig, 2012; Norris, 2006; Norris & Kinoshita, 2012; Norris & McQueen, 2008; Norris, Mcqueen, & Cutler, 2016; Reinisch, Jesse, & McQueen, 2010) has accumulated a wealth of evidence that the character of the cue-integration process is an approximately *Bayesian* assimilation of top-down beliefs and bottom-up evidence.

To offer an example specific to the intersection of phonetics, segmental phonology, and word recognition, this means that if I think I heard someone say ***sh****igarette*, I am likely to combine this acoustic information with my strong prior expectations about which words are (and are *not*) in the English lexicon and my rich experience as a speaker-hearer with

[32]Marr's *computational level of analysis* (Marr, 1982, Ch. 1) and the much older framework of *Ideal Observer Analysis* in psychophysics are important predecessors. See Geisler (2003) for an extended or Norris (2006, pp. 329-332) for an abbreviated review.

more or less likely production and perception errors to confidently infer that the speaker most likely said (or meant) to produce *cigarette* (Norris & McQueen, 2008).

With a framework of probability now introduced, I can now be formally transparent about exactly what notion of 'confusability' is used in existing synchronic-phonological literature and what the alternative is:

- If $\rtimes x_1, x_2, x_3 \ldots \ltimes$ is a segment sequence ('wordform') actually produced by a speaker, an *entirely* out of context measure of confusability[33] is (or is a function of) $p(\widehat{x}_i | X_i = x_i)$, the listener's beliefs about the speaker's intended *i*th segment given that the actual *i*th segment produced was x_i.
- A *local* measure of the confusability of x_i would be

$$p(\widehat{x}_i | x_{i-k} \ldots x_i \ldots x_{i+k})$$

  for $k \in \mathbb{Z}^+$ close to 1.
- A more psycholinguistically accurate measure of the confusability of x_i would be

$$p(\widehat{x}_i | G, x_1 \ldots x_{i-1}, x_i, x_{i+1} \ldots x_{i+k})$$

  i.e. the listener's beliefs about the speaker's intended *i*th segment given that
  - the actual current segment is x_i
  - the current wordform that x_i is part of was produced in some larger ('global') context G
  - the speaker has already produced acoustics for preceding segments $x_1 \ldots x_{i-1}$ (causes of perseveratory coarticulation and evidence that alters the listener's

[33] But one conveniently measured several times in varying degrees of completeness since the mid-20th century — see e.g. Miller and Nicely (1955), Wang and Bilger (1973), or Luce (1987).

top-down beliefs about what follows) in the wordform

- the upcoming segments (i.e. causes of anticipatory coarticulation) are $x_{i+1} \ldots x_k$.

2.3.2 Conclusion

In sum then, to identify the speech segment associated with acoustic data, it is empirically well established that listeners (as is generally the case in inference and perception in other domains of cognition) combine the 'raw' perceptual (acoustic) information associated with a speech sound that existing synchronic accounts consider with top-down expectations ('What word am I expecting the speaker to be producing?') that are incrementally updated as contextual cues unfold. These cues can in general be earlier segments in the word that have already unfolded, linguistic information at other levels of structure, or even non-linguistic information. Crucially, while existing synchronic-phonological accounts incorporate 'bottom up' acoustic confusability, they do not integrate it with any notion of incrementally adjusted top-down expectations.

While the research reviewed here indicates (as elaborated later in §2.5) that the architecture of existing synchronic-phonological accounts is inaccurate, mispredicts how perceivable any given speech sound type or token is, and underpredicts how much variation there should be in how perceptible a given segment is or what it's confusable with, it does not clearly spell out the systematic nature of such errors, where to look for them, or the magnitude of the errors. Accordingly, in the next section I derive and analyze a mathematical model of spoken word recognition (cf. Norris and McQueen 2008) to clarify and spell out the kinds of systematic errors in synchronic-phonological accounts make; in the section after, I then discuss an instantiation of this model with real psychoacoustic data on contextual perceptibility and use it to approximately measure the scope and magnitude of errors resulting from neglecting top-down expectations.

2.4 A mathematical model of spoken word recognition

In this section, I define a computational model of the process of spoken word recognition and then discuss its behavior and predictions with respect to variability in segmental perceptibility and the broader context of the chapter.

2.4.1 Model derivation

In this subsection, I lay out a Bayesian causal model of the word recognition process that spells out my assumptions about what affects what, and what simplifications I make in order to make use of available data. The key result is an expression for a listener's beliefs about what word is being produced given that a speaker has generated some acoustic data associated with a partial sequence of the segments in the word they are producing (Eq. (2.6)). This expression is used in the next subsection.

(2.4.21) a. The speaker chooses a single intended wordform $V = v^*$ — e.g. *cigarette* — from a set of words or word lemmas V, with the probability of choosing v^* given by a probability distribution $p(V)$.

b. The speaker determines something approximately like an intended segmental wordform $w^* = x_1^f = (x_1, \ldots, x_f)$ corresponding to their choice of v^*. In the running example, the speaker will determine that the segments [sɪgəɹɛt] correspond to the word *cigarette*. For ease of exposition, I will assume there is a unique correct segment sequence for a given v^* and will therefore identify v^* with w^*.

c. At any given point in time during this process, the speaker has completed producing an acoustic signal $a_0^t \sim p(A_0^t|\cdot)$ for some segmental prefix x_1^i of the current wordform — e.g. [sɪgə].

The listener's task is reasoning about the likely cause or explanation (actual intended wordform of the speaker W) of their observations:

$$p(W = x_1^f | a_0^t) \propto p(a_0^t | x_1^f) p(x_1^f) \tag{2.1}$$

A possible intended wordform is judged to have high probability given the acoustic data the more that it is likely to have been chosen by the speaker in the first place and the more that it is likely to have given rise to the observed acoustic data.

This is a model of a single listening event, and it directly references acoustic data that are both not the right abstraction in this chapter and that will be difficult to acquire and analyze at the relevant scale. Fortunately, the frequencies of a confusion matrix (Cutler, Weber, Smits, & Cooper, 2004; Luce, 1987; Miller & Nicely, 1955; Wang & Bilger, 1973)

$$p(Y_1^j \text{ segment sequence perceived} | X_1^i \text{ segment sequence underlying auditory stimuli})$$

refer to segments, are readily available, and can be interpreted as an aggregate measure of what kinds of percepts a given segment sequence is likely to give rise to. That is, they can roughly be thought of as the expected distribution over perceived segmental words given a produced segmental word, where expectation is with respect to possible acoustic signals given a produced prefix:

$$p(Y_1^j | X_1^i) \approx \int_{a_0^t} p(Y_1^j | a_0^t) p(a_0^t | X_1^i) \tag{2.2}$$

With less precision and less commitment, one may interpret Y_1^j as a discrete, approximate description of what a listener may perceive a speaker's produced acoustics as, and $p(Y_1^j | X_1^i)$ as a similar approximate description of $p(A_0^t | X_1^i)$. Accordingly, the reformulated version of

(2.1) is

$$p(W = x_1^f | y_1^i) \propto p(y_1^i | x_1^f) p(x_1^f) \tag{2.3}$$

$$= \sum_{x_{i+1}^f} \frac{p(y_1^i | x_1^i, x_{i+1}^f) p(x_1^i, x_{i+1}^f)}{p(y_1^i)} \tag{2.4}$$

$$= \sum_{x_{i+1}^f} \frac{p(y_1^i | x_1^i) p(x_1^f)}{\sum_{x_1'^f} p(y_1^i | x_1'^i) p(x_1'^f)} \tag{2.5}$$

assuming — for ease of exposition — that there is no effect of coarticulation and no insertion or deletion errors in perception. If, for example, a listener perceives y_1^f =e.g. [ʃɪgəɹɛt] (*shigarette*), their beliefs about the lexicon $p(X_1^f)$ will tell them both that this is not a segmental wordform in the lexicon, but that [sɪgəɹɛt] is. Their beliefs about the phonetics of their language $p(Y_1^f | X_1^f)$ tell them that x_j =[s] is a plausible segment to be misperceived as y_j =[ʃ]; together this suggests that a good explanation of their percept is the intended wordform x_1^f =[sɪgəɹɛt].

Eq. (2.3) allows us to measure how accurately the listener will be able to reconstruct the speaker's intended message given a specific perceived segmental prefix y_1^i. As mentioned earlier, an intended wordform x_1^{*f} with a so-far produced prefix x_1^{*i} may in general give rise to many different perceived wordforms y_1^i as a result of variation in production and noise in perception, and we want some aggregate measure that relates a speaker's intention to communicate $X_1^f = x_1^{*f}$ (while having produced x_1^{*i}) with the listener's beliefs about what X_1^f is — i.e. the listener's *expected* beliefs about X_1^f given that X_1^f is actually x_1^{*f} and

that x_1^{*i} has been produced so far:[34]

$$p(\widehat{X}_1^f = x_1^{'f} | X_1^i = x_1^{*i}) = \sum_{y_1^i} p(x_1^{'f} | y_1^i) p(y_1^i | x_1^{*i}) \tag{2.6}$$

$$= \sum_{y_1^i} \frac{p(y_1^i | x_1^{'i}) p(x_1^{'f})}{p(y_1^i)} p(y_1^i | x_1^{*i}) \tag{2.7}$$

$$= \sum_{y_1^i} \frac{p(y_1^i | x_1^{'i}) p(x_1^{'f})}{\sum_{x_1^f} p(y_1^i | x_1^i) p(x_1^f)} p(y_1^i | x_1^{*i}) \tag{2.8}$$

The $\widehat{\cdot}$ is used to suggest that the variable refers to the listener's *estimate* of the true intended wordform. Note that the left term in the product of (2.6) is Eq. (2.3).

In the next subsection, I discuss in more detail the incremental behavior of Eq. (2.6) as new segments are produced in terms of its component parts, and as facts about the lexicon change. I then locate this discussion in the larger context of the chapter.

2.4.2 Interaction of perceptibility and beliefs about the lexicon

Equipped with a mathematical model of word recognition in terms of Bayesian inference, we can analyze it and recover a crisp description of what kinds of systematically incorrect predictions about perceptibility result from neglecting the effect of incrementally adjusted top-down expectations, where within a language we might reasonably expect to find some good examples, and the predicted consequences for synchronic-phonological analyses.

According to Eq. (2.6), a typically perceptible prefix x_1^{*i} is one that *usually* gives rise (per $p(Y_1^i | x_1^{*i})$) to perceived prefixes y_1^i that are *usually* good enough evidence for the listener to assign a high degree of belief to the speaker's actual intended wordform (per

[34] Also note that exploration of this posterior is novel compared to other work on Bayesian models of word recognition — e.g. Norris and McQueen (2008) — but comparable to Eq. VII of Levy (2008b), a work on confusability at the sentence level.

$p(X_1^f|Y_1^i)$). Eq. (2.3) indicates how to peer inside $p(X_1^f|Y_1^i)$ and evaluate how good of evidence a *particular* channel string y_1^i is of any given intended wordform x_1^f: y_1^i is good evidence of x_1^f insofar as x_1^f is something the speaker is likely to have intended to say in the first place (per $p(X_1^f)$), and insofar as x_1^f is likely to have given rise to y_1^i, rather than other possible perceived prefixes (per $p(Y_1^i|\cdot)$). Insofar as the lexicon includes other words $x_1^{'f}$ with both high prior probability and that are good alternative explanations for observing y_1^i, the denominator will be higher and x_1^{*f} will be a worse explanation for y_1^i. Finally, note that the longer a produced prefix is, the more evidence the listener has about the speaker's intended goal and the fewer words there are at all that could explain the listener's percepts.

With respect to the broader context of the chapter, the distribution $p(Y_1^i|\cdot)$ exactly describes bottom-up sensory information — the context-free kind of facts about perceptibility that synchronic-phonological accounts hypothesize is directly present and referenced in phonological knowledge — where $p(X_1^f)$ describes top-down expectations. The interplay of these two terms described above makes the following predictions:

(2.4.22) a. **Prediction 1:** *Knowledge of the lexicon can **moderate** the effect of low perceptibility.* Tokens of segment types that are in general less perceptible (or that are in a phonotactic context where they are less perceptible than they otherwise would be) may occur in words that are on average relatively predictable, or (even in words that are typically not that contextually predictable) in an incremental context that makes the acoustically-confusable segment easy to correctly discriminate from alternatives.

b. **Prediction 2:** *Knowledge of the lexicon can **magnify and redirect** the effect of low perceptibility.* If the lexicon leads the listener to expect other subsequences much more so than the actual one and the actual one is less perceptible, then the subsequence will be even harder to accurately perceive, and what it's confusable

with will in general shift in the direction of what's predicted by lexicon-driven expectations.

c. **Prediction 3:** *The effect of knowledge of the lexicon should be strongest at extremes of word length.* Early in production of a word, acoustic evidence is relatively weak compared to top-down expectations; the local, acoustic-confusability-only model of perceptibility implicitly assumed by synchronic-phonological accounts would systematically over-estimate the perceptibility of segments occurring word-initially. As the speaker moves through producing a word, more and more evidence is revealed to the listener about what the word (likely) is and what the word (likely or even almost certainly) is not;[35] the local, acoustic-confusability-only model of perceptibility implicitly assumed by synchronic-phonological accounts would systematically under-estimate the perceptibility of segments occurring the further into a word they are.

   Similarly, segment tokens occurring in short incremental contexts could plausibly be caused by many possible total wordforms, but segments occurring in increasingly long incremental contexts could only be caused by one of decreasingly many long wordforms. In both cases, it is knowledge of what words are in the entire lexicon and how many have what lengths in the lexicon that underlies this effect.

I've described three salient trends about the systematic interaction of top-down expectations and bottom-up sensory cues. While I highlighted how this interaction should lead to systematic differences and more variation in the distribution of perceptibility than expected by synchronic-phonological accounts, below I spell out the corresponding predictions about the kinds of errors likely made in synchronic-phonological analyses of the phonology

[35]Indeed, this is a mathematical truth — conditioning must, in expectation, reduce expected surprisal: $H(B|A) = \mathbb{E}_a\, H(B|a)$ must be no greater than $H(B)$, with equality holding iff A and B are independent.

of individual languages. I note first that existing synchronic-phonological accounts (like most others in phonology) have never been tested on a dataset consisting of most or all of an entire language's lexicon; a large part of the reason why is that there is at present no lexicon which (human) analysts have annotated every (or most) wordforms with an underlying representation, nor is there currently a general, practical, and accurate algorithm for doing so at scale. As a result, the most that can be done in the meantime is to predict what would happen assuming future work (building on e.g. Pater, Jesney, & Smith, 2012) provides a practical and effective algorithm for inferring underlying representations at the scale of entire lexicons. Second, recall from e.g. §2.2.4.2.2 that synchronic-phonological accounts translate statements about the relative perceptibility of a contrast directly into faithfulness constraints and their relative ranking: a contrast that is more perceptible in one context is more important to preserve than the same contrast in less perceptible contexts.

The general character of errors synchronic-phonological accounts of perceptibility effects would make are wrong predictions about what accounts for the phonotactics of a language or the need to posit constraints or constraint rankings otherwise unmotivated in order to patch up incorrect predictions made by using perceptibility constraints based on strictly local ('out-of-(global)-context') confusability measures. The three most salient cases are:

(2.4.23) a. **Out-of-context confusable but in-context not-confusable contrasts.** The faithfulness constraints for such contrasts will be ranked lower than their actual perceptibility warrants. Such contrasts would occur in contexts where top-down expectations and knowledge of the lexicon moderate the effects of low out-of-context perceptibility.

b. **Out-of-context confusable and in-context even-more confusable contrasts.** The faithfulness constraints for such contrasts will be ranked higher than their

actual perceptibility warrants. Such contrasts would occur in contexts where top-down expectations and knowledge of the lexicon magnify the effects of low out-of-context perceptibility.

c. **Out-of-context not-confusable but in-context confusable contrasts.** The faithfulness constraints for such contrasts will be ranked higher than their actual perceptibility warrants. Such contrasts would occur in contexts where top-down expectations and knowledge of the lexicon overwhelms the effects of moderate or possibly even relatively high out-of-context perceptibility.

In this section, I

(2.4.24)
a. formalized the empirical results described in the previous section as a computational psycholinguistic model,
b. described the behavior of the model,
c. described the ways in which the perceptibility of tokens of a given segment type will systematically vary based on the structure of the entire lexicon and the epistemic context it occurs in, and
d. described corresponding errors in the phonological descriptions synchronic-phonological accounts should be making.

In the following sections, I will describe instantiating the word recognition model using real data and then empirically showing the difference between perceptibility in-context (i.e. in a real lexicon) vs. out-of-context.

2.5 Variation in the perceptibility of the American English inventory

In this section, I use a transcribed lexicon of American English, corpus-derived frequency estimates (Davies, n.d.), and psychoacoustic data (Warner et al., 2014) to construct an approximation to the model outlined previously. Crucially, the psychoacoustic data allows for a limited model of the effects of phonotactic context (i.e. coarticulation) on confusability — a triphone-to-uniphone channel model $p(y_i|x_{i-1}^{i+1})$.

To show the variation in perceptibility caused by context and the magnitude of this effect, I construct an artificial lexicon from the same data consisting only of licit word-internal triphones found in the natural lexicon and place a uniform distribution on it. By construction, the only ways top-down expectations affect 'word' recognition in this setting are via the size of the set of triphones and categorical phonotactics of the natural lexicon: the effect of phonotactic context on confusability via coarticulation (generally, the confusability term of Eq. (2.6)) is about as strong as it could be, and the lexicon term is about as weak as it could be.

Equipped with these two lexicons, I use information-theoretic measures to show aggregate differences in perceptibility for individual segment types between the two lexicons — viz. that there is much more variation in perceptibility across the 'global' contexts of the natural lexicon than the almost purely local contexts of the artificial lexicon, and that the magnitude of the effect of global context on perceptibility is large.

Below I describe the construction of the approximate model, explain the choice of information-theoretic measures and aggregate visualizations, and then present and describe the results of applying the measures to the two lexicons.

2.5.1 Constructing an approximate word recognition model

2.5.1.1 Diphone gating data

The model of segmental confusability presented here is ultimately based on the diphone gating experiment data of Warner et al. (2014). Participants listened to gated intervals of every phonotactically licit diphone of (western) American English and attempted to identify the full diphone they thought was being produced during the interval. Along with earlier work by some of the same researchers on Dutch (Smits, Warner, McQueen, & Cutler, 2003; Warner, Smits, McQueen, & Cutler, 2005), this represents by far the richest and most comprehensive acoustic confusion matrix data of its kind, and the only one capable of also offering a relatively comprehensive window into the perceptual effects of local phonotactic context due to coarticulation. This means it is uniquely well-suited for comparing what the facts of relative perceptibility in an entire language with vs. without a more comprehensive model of perceptibility than that explored by synchronic-phonological accounts of perceptibility effects.

To construct the set of stimuli diphones, Warner et al. identified all adjacent pairs of segments within and between words based on an electronic pronouncing dictionary of about 20,000 American English wordforms. A set of approximately 2,000 phonotactically licit diphones were extracted from this transcribed lexicon. At least one stimulus nonsense word was created per diphone by inserting the diphone into a small phonotactic environment systematically chosen to avoid word edge effects on pronunciation and acoustics, to aid pronounceability (in light of phonotactic effects on stress and syllable structure), and to avoid predictability of the diphone from the preceding context.[36]

Each nonsense word was produced by a phonetically-trained speaker who was

[36]This also means that the channel model will be a less accurate representation of confusability near word-edges.

monolingual until she was a teenager and whose native dialect was well-matched to listeners. A recording of each stimulus wordform was then marked up with (generally) six temporal gates: the first about a third of the way through the first segment, the second about two-thirds of the way through the first segment, the third all the way through the first segment, the fourth a third of the way through the second segment, etc. For each stimulus wordform, one recording was created for each gate, starting at the beginning of the original recording and going all the way up to a gate location, followed by a ramping procedure (rather than truncation or white noise) to avoid systematically biasing confusion data.

Twenty-two students at the University of Arizona completed the study and contributed to the confusion data used here. Each study participant listened to and attempted to identify stimuli recordings in a randomized order over the course of about 30 one-hour sessions following an initial instruction and practice period. In each trial, participants heard a gated stimulus recording seated at a computer.[37] If the recording included a preceding context, this context was displayed on the screen. The participant then selected the stimulus diphone they thought was in the recording (i.e. not including context). Contexts and response segments were presented using English grapheme sequences based on typical English spellings; competence in understanding these symbols and combinations was part of the practice period.

From this response data, each gate of each stimulus diphone can be associated with a frequency distribution over response diphones. Only the response data for gates corresponding to the end of each segment of the diphone were used in the current study. Principally for reasons of data sparsity, the distinction in the gating data between stressed and unstressed versions of each vowel were collapsed. Because the transcription lexicon lacked taps, alveolar taps were merged into [t].

[37] See Grosjean (1980) for reference on the gating paradigm.

2.5.1.2 Language model

The model of the previous section references a prior probability distribution over segmental wordforms. Given the goals of this chapter, a unigram language model is complex enough to serve the goals of the chapter and actually means that the measures reported here will systematically *under*estimate the true effects of context on perceptibility.

Orthographic word frequencies were taken from the top 1,000,000 words in the Corpus of Contemporary American English (Davies, n.d.). The complete corpus ('COCA') contains about half a billion words, collected from text across a wide variety of genres over the last 30 years.

2.5.1.3 Transcribed lexicon

For transcriptions, the same dictionary was used as Warner et al.. Warner et al. do not indicate where the transcriptions came from beyond providing a URL. The transcriptions in that file appear to be the same as those used in Sejnowski and Rosenberg (1987), although other annotations differ slightly. Sejnowski and Rosenberg note that the transcriptions come from an an unspecified edition of *Merriam-Webster's Pocket Dictionary*. There are about $20,000$ orthographic words in the dictionary and each is uniquely associated with a segmental transcription.

Words were dropped from this lexicon if

(2.5.25) a. They contained triphones that could not be modeled by the channel model.

b. They were not in the language model.

c. Their normalized frequency after applying the previous two exclusion criteria was below 5×10^{-7}.

This resulted in a final lexicon with about $12,000$ words.

The Western variety of American English native to the speaker and listeners of

the gating data has undergone the *cot-caught* merger; transcriptions in this lexicon were aligned accordingly.

2.5.1.4 Channel model

The *channel model* describes the conditional distribution $p(Y_1^f | X_1^f)$ over what sequence of segments y_1^f a listener will perceive (e.g. [ʃɪgəɹɛt], *shigarette*) given the full intended sequence x_1^f (e.g. [sɪgəɹɛt], *cigarette*). We estimate this distribution using the diphone gating data in §2.5.1.1. I make the simplifying assumption that the channel distribution for segment y_i is conditionally independent of all other y_j ($j \neq i$) given intended segments x_{i-1}, x_i, x_{i+1}.

By conditioning on adjacent segments, some effects of coarticulation on confusability are captured. For example, recall that nasals before oral stops are systematically likely to be misheard as having the same place of articulation as the stop: x_1^f =[ɑnpɑ] (alveolar nasal before labial stop) is more likely to be misperceived as y_1^f =[ɑmpɑ] (a labial nasal) than the reverse, and a confusion of [n] for [m] is comparatively less likely when [n] is between vowels as in [ɑnɑ] (J. J. Ohala, 1990a).

For each gate $g \in \{3, 6\}$ and for each diphone $x_1 x_2$, the response data from §2.5.1.1 induce a conditional frequency distribution over channel diphones $f_g(y_1, y_2 | x_1, x_2)$. These frequency distributions were smoothed by adding a pseudocount (0.01) to every channel diphone in every distribution; the distributions were then normalized to define a smoothed pair of diphone-to-diphone channel distributions $p_g(y_1, y_2 | x_1, x_2)$. From the marginals of these distributions an approximation (Eq. 2.9) of the triphone-to-uniphone channel distribution can be defined via their geometric mean:[38]

$$p(y_i | x_{i-1}, x_i, x_{i+1}) \propto \sqrt{p_3(y_i | x_{i-1}, x_i) p_6(y_i | x_i, x_{i+1})} \tag{2.9}$$

[38] A full triphone-to-triphone channel distribution was not used for reasons of tractability.

As a compact summary of Fig. 2.3 shows the negative log of this marginalized over contexts, assuming a uniform distribution over contexts. The pseudocount level was chosen experimentally by choosing among a few pseudocount levels to find one whose uniphone error probability (averaged over all segment types) is close to that of the marginal uniphone error probability obtained from the diphone gating distributions.

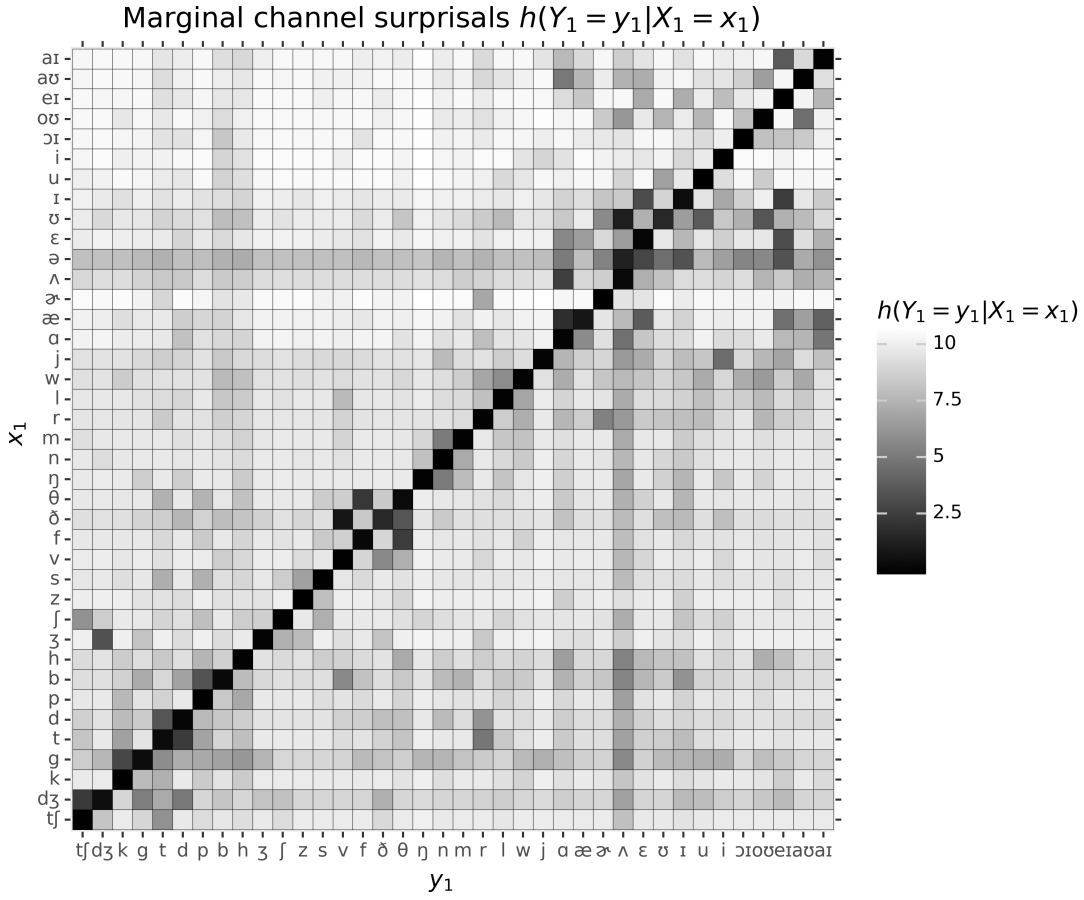


Figure 2.3: The surprisal (in bits) of Y_1 given X_1, marginalizing over phonotactic contexts X_0, X_1, assuming a uniform distribution on $X_0 \times X_1 \times X_2$.

With the simplifying assumption that only substitution errors are possible,[39] we

[39]The gating data does not provide information for estimating the probability of deletion or insertion errors.

obtain a string-to-string channel model:

$$p(y_1^i|x_1^i) = \prod_{j=1}^{j=i} p(y_j|x_{j-1}, x_j, x_{j+1}) \tag{2.10}$$

Note that the most similar previous channel model (Norris & McQueen, 2008) was based on Dutch gating data (Smits et al., 2003) comparable to that used here. Norris and McQueen did not construct a triphone-to-uniphone channel model, but made use of all gates and also allowed investigation of word boundary identification — i.e. word boundaries were not considered as given, and they instead ultimately defined a channel model on (in general) multi-word segment sequences.

2.5.1.5 Approximate model

The entire distribution $p(\widehat{X}_1^f|X_1^{i+1})$ is impractical to calculate in its entirety given its size, largely due to the number of channel strings for any given length that must be summed over in the normalization term of Eq. (2.12). Fortunately, because each segment type is actually only relatively confusable with a small number of other segment types, and the probability of any one recognition error is generally small, most prefixes are only confusable with a small number of other prefixes. Accordingly, $p(\widehat{X}_1^f|X_1^{i+1})$ can be approximated by only doing calculations for wordforms and prefixes that are within a small edit distance of each other, as elaborated below.

(2.5.26) a. Let $D_H(u, v)$ denote the Hamming distance[40] between two strings u, v.

b. Let x_1^f be a wordform and $x_1'^i$ be a prefix of an arbitrary wordform in the lexicon.

c. If x_1^f has any prefixes r such that $D_H(r, x_1'^i) \leq k$, then we say that x_1^f and $x_1'^i$ are *k-cousins*.

[40]Hamming distance is the number of symbol substitutions it takes to transform one string into another if they are the same length; otherwise the distance is infinite.

Intuitively, if a wordform x_1^f and a prefix $x_1'^i$ are only k-cousins for large k, then x_1^f is almost certainly a word where $p(\widehat{X}_1^f = x_1^f | X_1^{i+1} = x_1'^i)$ is extremely close to 0 and difficult to numerically calculate with any precision anyway. Accordingly, $p(\widehat{X}_1^f = x_1^f | X_1^{i+1} = x_1'^i)$ was only calculated for (wordform, prefix) pairs that were 2-cousins; the resulting distribution was then normalized.

2.5.2 Analysis

The goal of this subsection is to precisely define some ways to measure the effects of all contexts in a given lexicon on the perceptibility of each segment type. These measures will then be compared for the actual English lexicon and the artificial triphone lexicon constructed as described at the top of §2.5.

2.5.2.1 Information measures and their interpretation

Here I briefly introduce the less common information measures used in this chapter, focusing on their interpretation and their relevance. Given its prevalence in the last two decades of language research, I assume the reader is familiar with the interpretation of the *surprisal* of the outcome of a random variable $h(A = a) = -\log p(a)$, the *entropy* of a random variable $H(A) = \sum_a p(a)h(a)$, and their conditional variants $h(a|b)$, $H(A|B)$. See e.g. J. T. Hale (2003, 2006) or Levy (2005, 2008b) for introduction in the context of psycholinguistics, Stone (2015) or Cover and Thomas (2012) for elementary or advanced textbooks, and Csiszár (2008) or Ince (2017) for satisfying interpretations, descriptions of the structure, and ways of motivating classic information measures.

The *mutual information* between two discrete random variables A and B can be thought of impressionistically as a discrete analogue of correlation. More precisely, it tells you how much, on average, knowing the value of one variable reduces your uncertainty

about the other:

$$I(A; B) = H(A) - H(A|B)$$

The *pointwise mutual information* between two outcomes similarly describes how knowing that one variable has a specific value reduces your surprisal that the outcome of the other variable has a specific value:

$$i(a; b) = h(a) - h(a|b)$$

I will use $i(\cdot;\cdot)$ to describe how the speaker's production of incremental context changes the listener's surprisal of the actual ith segment.

To compare two probability distributions p, p' over the same event space A, the *Kullback-Leibler divergence* is commonly used:

$$D_{KL}(p||p') = H(p, p') - H(p)$$

where the notation $H(p)$ indicates $H(A)$ with p is used as the distribution for A, and $H(p, p')$ denotes the *cross-entropy* of p and p':

$$H(p, p') = \sum_a p(a) \log \frac{1}{p'(a)}$$

The cross-entropy can be thought of as the expected surprisal you would experience if you thought A was distributed according to p' when it is actually distributed according to p. The Kullback-Leibler (KL) divergence is then the expected *excess* surprisal you will experience relative to someone who knows the true distribution. Accordingly, it will be 0 iff p and p' are exactly the same and increase the more distinct p is from p'.

The *Jensen-Shannon* divergence is (roughly) a symmetric variation of KL di-

vergence. To understand its interpretation, consider the following scenario: suppose A is distributed according to m, a binary mixture of p and p':

$$m(a) = \lambda p(a) + (1 - \lambda)p'(a)$$

That is, to sample from m, you first flip a weighted coin C that with probability λ comes up heads. If it comes up heads, you sample from p, and otherwise you sample from p'. Suppose you flip this coin — *without* showing me the outcome — and it comes up heads. My excess surprisal relative to yours is given by

$$D_{KL}(p||m)$$

On average, if we repeat this process, my expected excess surprisal relative to yours is given by the λ-divergence of p from p':

$$D_\lambda(p||p') = \lambda D_{KL}(p||m) + (1 - \lambda)D_{KL}(p'||m)$$

Crucially for its interpretation, it can be shown that

$$D_\lambda(p||p') = I(A; C)$$

That is, the λ-divergence of p from p' is the average information gained about A by knowing which distribution it will be sampled from, and (because mutual information is symmetric) equivalently the information gained about the coin flip C by observing A, on average. The Jensen-Shannon (JS) divergence of p from p' is exactly this for $\lambda = \frac{1}{2}$ — a fair coin flip:

$$D_{JS}(p, p') = \frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(p'||m)$$

When p and p' are very similar, knowing the outcome of A tells me very little about the outcome of the hidden coin flip, and so $D_{JS}(p, p')$ is low. In the context of this chapter, a rough description of the main usage of JS divergence is that the 'coin' C will control which of two segment types x^*, x' the speaker actually produced, and the observed event A will be what the listener thinks the speaker actually produced.[41] The JS divergence, then, will indicate how distinct x^* is from x', on average: when the distributions are very similar, the JS divergence will be close to zero.

For the sake of illustration of JS divergence, compare Figures 2.4 and 2.3. Where 2.3 directly shows (log-transformed) marginal channel distributions $p(Y|X)$, Figure 2.4 shows

$$\overline{D}(x^*, x') = D_{JS}(p(Y|X = x^*)||p(Y|, X = x')) \tag{2.11}$$

for each (x^*, x') pair. As a summary of the overall similarity of pairs of distributions (rows) from Figure 2.3, Figure 2.4 unsurprisingly shows less variation and has most ink concentrated near the diagonal.

2.5.2.2 Operationalizing segmental perceptibility

With respect to the model described in §2.4.1, the mathematical object containing all information about the perceptibility of a segment *token* of x^* in some intended word token x_1^{*f} given that x_1^{*i} has been produced is given by marginalizing Eq. (2.6) over the listener's expected beliefs about $\widehat{X}_1^{i-1}$ and $\widehat{X}_{i+1}^{f}$:[42]

$$p(\widehat{X}_i = x'|X_1^i = x_1^{*i+1}) = \sum_{x_1'^{i-1}, x_{i+1}'^{f}} p(\widehat{X}_1^f = x_1'^f|X_1^i = x_1^{*i+1}) \tag{2.12}$$

[41] As clarified later, I will be comparing pairs of listener posteriors, *not* pairs of channel distributions.
[42] Note that this equation has also been adjusted slightly to reflect coarticulation.

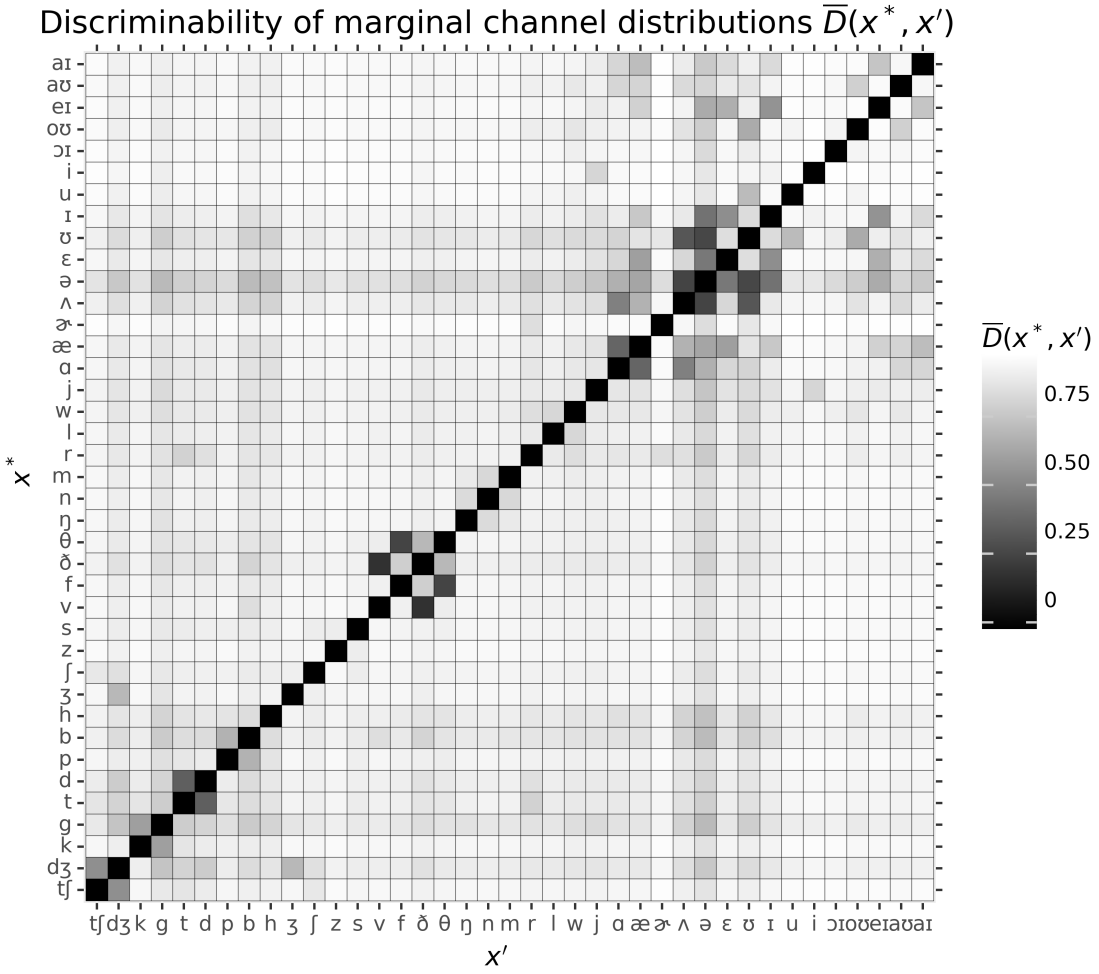


Figure 2.4: Similarity of marginal channel distributions for each pair of segment types x^*, x'.

For clarity below, I will refer to the produced incremental context X_1^i and upcoming segment X_{i+1} together as the joint random variable C: i.e. $C = X_1^i \times X_{i+1}$. With respect to Eq. (2.12), there are two natural places to look:

(2.5.27) a. The probability the listener assigns to the speaker's ith actual intended segment type $p(\widehat{X}_i = x^* | X_1^i = x_1^{*i+1})$.

b. The listener's entire expected distribution about the ith segment $p(\widehat{X}_i | X_1^i = x_1^{*i+1})$.

I will do both. To analyze a segment *type* x^*, I will examine these two objects for all tokens of x^* in the lexicon.

I begin by focusing on 2.5.27a. A simple first measure to compare between the artificial and the natural lexicons is the relative average confusability of segments in one vs. the other:

$$h(\widehat{X} = x^* | X = x^*) \tag{2.13}$$

$$h(\widehat{X} = x^* | X = x^*, c) \tag{2.14}$$

Eq. (2.13) measures how well the listener has recovered the speaker's intended segment, marginalizing over contexts. Eq. (2.14) shows this in a particular incremental context $c = x_1^{*i-1}, x_{i+1}^*$. Because the channel model is the same between the artificial and natural lexicon, clear differences in these measure for a given segment will be due to the presence of stronger top-down expectations, informative incremental contexts, and structured variation in the natural lexicon.

To measure how a specific incremental context $c = x_1^{*i-1}, x_{i+1}^*$ changes a listener's expected beliefs that $\widehat{X}_i = x^*$, the appropriate tool is *pointwise mutual information*:

$$i(\widehat{X}_i = x^*; C = c | X_i = x_i^*) = h(\widehat{X} = x^* | X = x_i^*) - h(\widehat{X}_i = x^* | X_i = x^*, C = c) \tag{2.15}$$

In words: this is the *change in surprisal* that $\widehat{X}_i$ is x^* caused by the specific context c, relative to the expected surprisal that $\widehat{X}_i = x^*$ averaged over all contexts in the lexicon that x^* occurs in.

This average is nontrivial to correctly define for the natural lexicon due to variation in length. To understand exactly what this average means and how it is calculated, it is

clearest to first define a joint distribution between total contexts (all segments to the left of some segment within a word, plus all segments to the right within a word) and segment types. A *context token* is specified by a choice of word w and a position i within the word. The full joint distribution is then $p(X = x, I = i, W = w)$, specified by the following generative process:

(2.5.28) a. Sample a word w according to the unigram language model prior $p(W)$.

b. Sample an index i from a uniform distribution on indices $[1, |w|]$. All other index values have probability 0.

c. $p(x|i, w) = 1$ iff the ith segment of w is in fact of type x and 0 otherwise.

I.e.

$$p(x, i, w) = p(x|i, w)p(i|w)p(w) \tag{2.16}$$

The distribution on total contexts conditioned on a particular segment type x^* is then exactly given by $p(I, W|x^*)$.

In the case of the probability distribution underlying $h(\widehat{X} = x^*|X = x^*)$,[43] the relevant context model refers to incremental contexts — essentially prefixes. The marginal distribution on prefixes $p(R)$ has exactly the same structure and generative process as the marginal distribution $p(I)$ in 2.5.28: there is a joint distribution on words and prefixes

$$p(R = r, W = w) = p(r|w)p(w) \tag{2.17}$$

where every licit prefix of a given word w has uniform probability. Given the marginal

[43]The subscript i in Eq. (2.15) references position and ultimately segment identity with respect to the particular context c and is only meaningful in the full scope of that equation — not within the definition of the average contextual surprisal of a segment type.

distribution $p(R)$, then, there is a joint distribution

$$p(X_{-2} = x, R = r) = p(x|r)p(r) \tag{2.18}$$

where $p(x|r)$ is 1 iff the second-to-last segment of prefix r is of type x and 0 otherwise. Let C, as earlier, denote the set of incremental contexts — the set of lexical prefixes c with a 'hole' in the second-to-last position. Then,

$$p(\widehat{X} = x^*|X = x^*) = \sum_c p(\widehat{X} = x^*|X = x^*, C = c)p(C = c|X = x^*) \tag{2.19}$$

The rightmost term is calculated as described above, and if $c = x_1^{'i-1}, x_{i+1}^{'}$ for some i, then we have:

$$p(\widehat{X} = x^*|X = x^*, C = c) = p(\widehat{X} = x^*|X_1^{i-1} = x_1^{'-1}, X_i = x^*, X_{i+1} = x_{i+1}^{'}) \tag{2.20}$$

I.e. the left term inside the summation of is nothing more than an incremental posterior calculation, as one would expect.

Turning to 2.5.27b, we can compare the confusability of one segment type x^* is to that of another x' in a particular context $c = x_1^{i-1}, x_{i+1}$ that they can both occur in via what I will denote as $D_{\bar{c}}(x^*, x')$:

$$D_{\bar{c}}(x^*, x') = D_{JS}(p(\widehat{X}_i|X_i = x^*, c)||p(\widehat{X}_i|X_i = x^{'}, c)) \tag{2.21}$$

Note that this is symmetric because JS divergence is symmetric.

To measure how similar this measure of confusability is across all common contexts $C(x^*, x')$, we can take an expectation over common contexts

$$\widehat{D}_{\bar{c}}(x^*, x') = \sum_c p(c|X = x^*, c \in C(x^*, x'))D_{\bar{c}}(x^*, x') \tag{2.22}$$

Note that the choice of expectation distribution (conditioning on x^*) means $\widehat{D}_{\bar{c}}(x^*, x')$ is *not* symmetric; this permits separate comparison of the typical confusability of x^* and x' in contexts typical for x^* vs. typical for x': x' might be similar in its confusability to x^* in the contexts that x' typically occurs in, but that does not entail the reverse must be the case.

2.5.2.3 Results

Figures 2.5-2.6 show Eq. (2.13) in the artificial and natural lexicons (respectively), summarizing for each segment type how likely it is to be confused, on average, in the contexts it occurs in. As the axes indicate, the general effect of incremental top-down expectations is to make segments less confusable. Comparing the two graphs also shows that while some trends in the relative ordering of confusability in the artificial lexicon are roughly preserved in the natural lexicon (e.g. within glides and liquids, within nasals, and within fricatives), the preservation is only approximate, and there is no shortage of other differences — incremental context and top-down expectations clearly affect which segments are on average relatively confusable, supporting the key claim of this chapter that out-of-context measures of confusability do not, in general, reflect in-context confusability.

Turning to Eq. (2.14) and examining variation in more detail, Figures 2.7 and 2.8 again show the generally depressing effect of natural incremental context on confusability, but also, intriguingly, reveal that many segments — including those whose confusability in the artificial lexicon is low and relatively stable — have a long tail of incremental contexts where they as confusable or more so than they typically are in the artificial lexicon. Figures 2.9-2.10 show how this relates to position within the word.[44] They indicate that in the limit of increasing within-word position, incremental top-down expectations and knowledge of the lexicon will typically leave a listener with almost no uncertainty about segment identity and nearly complete confidence in the speaker's actual intended segment type. Per §2.5.2.1,

[44]Datapoints associated with word-initial and word-final contexts are omitted because the channel model is in general a relatively inaccurate approximation there.

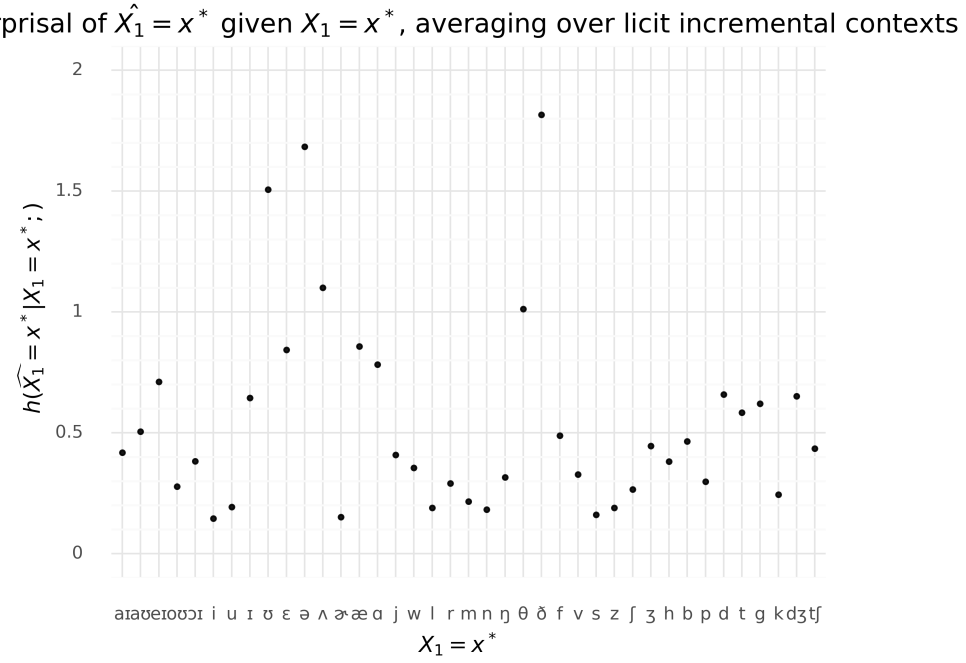


Figure 2.5: Posterior contextual surprisal of each segment type x^* in the artificial lexicon, marginalizing over contexts.

the difference between the first pair of plots and the second can be understood via pointwise mutual information. The next set of plots show the role of incremental context in creating the differences just discussed.

Figures 2.11/2.12 and 2.13/2.14 are matched pairs of graphs showing (for the artificial lexicon and natural lexicon, respectively) the values of Eq. (2.15), grouped by each segment: each column is a segment type x^*, and each datapoint in a given column corresponds to a calculation of Eq. (2.15) for an incremental context. Figures 2.11/2.12 show the most detail but omit some data, where 2.13/2.14 show less detail but 2.14 still omits some data; Figure 2.15 shows the full distribution for the natural lexicon. Figures 2.16-2.17 show how the effect of context on a given segment token varies as function of that segment's distance from the left edge of the word.

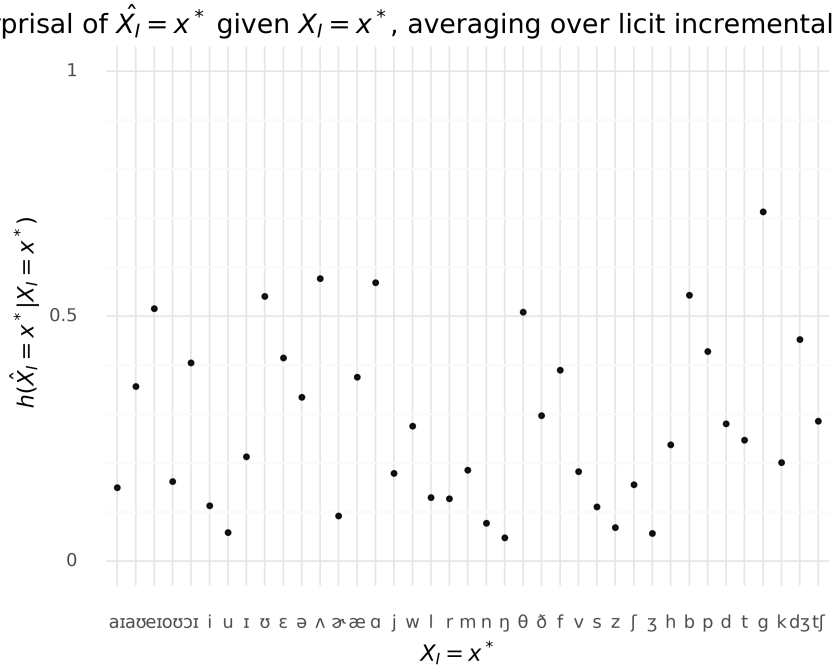


Figure 2.6: Posterior contextual surprisal of each segment type x^* in the natural lexicon, marginalizing over contexts.

These figures show several trends: The effect of context (via phonotactics and coarticulation) on perceptibility is weaker in the artificial lexicon, with some segments (e.g. [i,u,ɚ,l,n,s,k]) showing a consistent effect of context of about 0 bits, with others showing a range of the effect of local context of about 1 bit. In contrast, in the natural lexicon, the typical effect of incremental context is typically positive (if small) across all segments, including for segments whose out-of-context perceptibility is relatively higher. This supports Prediction 1 (2.4.22a) of §2.4.2. A the same time, again visible is the long-tail of misinformative contexts most segment types x^* in the natural lexicon have — many contexts where x^* is typically *more likely* to be misperceived than it otherwise would be, and much more likely to be misperceived than one would expect given only a model of local context effects comparable to those displayed in the artificial lexicon. This supports

Prediction 2 (2.4.22b) of §2.4.2. As Figures 2.16-2.17 show, these segment tokens are principally word-initial or in relatively short words. The clear effect of position supports Prediction 3 (2.4.22c) of §2.4.2.

Turning to aggregate patterns of what segments are typically confusable with what other segments, Figures 2.18 and 2.19 show (for the artificial and natural lexicons, respectively), for each pair of segments x^* and x' the value of Eq. (2.22), $\hat{D}_{\bar{c}}(x^*, x')$.

While the artificial lexicon displays some clear and interpretable patterns (vowel-vowel similarity, similarity between consonants differing in one of manner or place) of which segments have contextually similar confusion distributions, on the whole segments have relatively distinct marginal posteriors in the artificial lexicon. In contrast, though some of the trends of similarity in the artificial lexicon carry over into the natural one, in general there is much more variation and marginal posteriors are much less pairwise distinctive. This supports Prediction 2 (2.4.22b) of §2.4.2.

2.6 Discussion

2.6.1 Implications for synchronic-phonological accounts

Perceptibility effects relate robustly observed trends under experimental conditions in what speech sounds are confusable with which others — sometimes as conditioned by particular local phonotactic contexts — with trends in common synchronic and diachronic patterns in phonology. Recall that diachronic-phonetic accounts of the typological prevalence of these patterns broadly point to processes of perception, production, learning and phonologization to explain this. Contemporary synchronic-phonological accounts (e.g. Steriade 2001b, 2008; Wilson 2006), in contrast, do not dispute this line of inquiry, but instead propose that this typological trend in phonological grammars could be explained by the direct representation of phonetic facts about relative perceptibility and perceptual similarity

and the sensitivity of these two phenomena to relatively local phonotactic context in grammatical representations. (Recall the *direct translation* of facts about relative perceptibility into faithfulness constraints described in §2.2.4.2.2.) Some accounts go further and propose the hypothesis that this domain-specific knowledge is innate and that the universality of innateness explains the typological distribution and prevalence of perceptibility effects (Wilson, 2006).

The results of the previous section show that the approximate perceptibility of tokens of a given segment type

(2.6.29) a. Varies greatly.

b. Varies not only as a function of coarticulation with local phonotactic context, but also as a function of global token context and its epistemic effects on perception, well above and beyond what is predicted purely by focusing on coarticulation and local context.

c. Varies as a function of the structure of the entire lexicon and gradient, context-sensitive, and volatile facts about the language and the contexts of use like the relative probability of different wordforms.

Below I elaborate on why this means contemporary phonological accounts cannot achieve their explanatory goals as currently formulated and why revising them (especially nativist ones) to accurately model confusability would make them far less plausible.

2.6.1.1 Revision would be self-consistent

Almost universally, synchronic-phonological accounts of perceptibility effects reference the actual acoustic facts or acoustic experiences of listeners as the cause and/or intensional basis for defining phonological constraints. Accordingly, I first argue that it would only be self-consistent for synchronic-phonological accounts to be revised in light of

the results of the previous section — even setting aside the analytical errors and descriptive inadequacy §2.4.2 predicts of synchronic-phonological accounts.

Within contemporary synchronic-phonological accounts,[45] Steriade (2001b)'s original proposal references *perceptual experience* as the basis of the P-map. Given an operationalization of 'similarity' judgements in terms of confusability (as in e.g. Wilson 2006) and descriptive psychological evidence of how human perception *actually* functions, using perceptibility measures that incorporate the effect of top-down expectations and context amount to a much more accurate model of human perceptual experiences. To date there has been no explicit commitment in synchronic-phonological literature to specifically local measures of confusability; their consistent use and reference since early phonetically-based phonology work simply reflected what was familiar to researchers at the time. It seems uncontroversial, then, to assume that updating synchronic-phonological accounts to reflect what perceptibility is actually like is reasonable.

2.6.1.2 A revised architecture would be implausible

In spite of my argument that synchronic-phonological accounts would be more consistent with their own stated aims if they used a psycholinguistically accurate notion of confusability incorporating global context and expectations, I argue that it is also the case that revising these theories to use an accurate inference process would demand a number of architectural changes whose result would likely be deemed undesirable.

First, in a revised synchronic-phonological theory, constraints would have to be able to reference arbitrary stable cues – i.e. cues at arbitrary parts of the linguistic hierarchy or ones that are altogether non-linguistic. My modeled task of isolated wordform recognition is incredibly simple and therefore only requires reference to incremental prefixes, but as noted before in §2.3, there is robust psycholinguistic evidence that listeners integrate a wide

[45] I.e. similarity-based ones.

variety of linguistic cues at all levels of the linguistic hierarchy as well as non-linguistic information in the course of recognition and comprehension. Much phonological work is concerned with *limiting* the interaction of a single other level of the linguistic hierarchy with mechanisms of phonological theory (e.g. Keating, 1985): the possibility that cues from *every* level of the linguistic hierarchy or even altogether non-linguistic cues could require representation would be problematic from such a perspective.

A consequence of this is that the number of (combinations of) cues that affect a confusability calculation is plausibly vastly greater than appreciated: to accurately capture perceptibility effects requires either a combinatorically horrific explosion in the number of possible constraints or an ingenious (and heretofore unspecified) learning and/or evolutionary mechanism (in the case of substantively-biased phonology) for navigating this immense space of representations.

Third, a revised phonological grammar would also require reference to probabilities (or an approximation thereof) in order to correctly calculate confusability. As well, because changes to one wordform can in general affect probabilities elsewhere, a consequence of this is that the phonological grammar of an individual would require simultaneous optimization of the *entire* lexicon.[46]

Given that

- few other phonological phenomena seem to require anywhere near as many constraints or simultaneous optimization for an acceptable analysis
- there is still no compelling story as to why phonetic facts *need* to be represented in constraint-based phonological representations to account for perceptibility effects

revising existing synchronic-phonological theories seems unlikely to be compelling to phonologists working in constraint-based theories, relative to diachronic-phonetic accounts

[46]This echoes communicatively-oriented constraint-based synchronic-phonological work (see e.g. Padgett 1997, 2003 or Łubowicz 2003).

of perceptibility effects specifically and simpler architectures for constraint-based phonology. Finally, given the variation across languages and over time in how frequent (and how informative) different segment types and contexts are (Cohen Priva, 2012), the domain-specific representational knowledge of relative perceptibility posited by substantively-biased phonology are especially implausible absent an explanation of why it would instead reflect only cross-linguistically stable acoustic facts of perceptibility and how such acoustic facts could ever come to be genetically specified.

2.6.2 Implications for diachronic-phonetic accounts

Although diachronic-phonetic accounts are not the main focus of this chapter, the empirical results reported here also have tentative ramifications for them — beyond offering additional evidence that they are a better direction for explaining perceptibility effects than constraint-based synchronic-phonological accounts.

As stated earlier, existing diachronic-phonetic accounts have been particularly strong at connecting laboratory results on the effects of particular local phonotactic contexts on confusion and perception to common phonological patterns. In contrast, no work to my knowledge has attempted a detailed model or theory of why perceptibility effects arise in the languages they do or when they do: if two languages have comparable segment inventories and comparable phonotactics, why should only one end up with certain perceptibility effects instead of both? For any particular language that has a perceptibility effect or is in the process of phonologization of one, what words in the lexicon are most likely to give rise to the effect? Assuming that the gating data and channel model presented here are decently representative of confusability in naturalistic speech, the measures and figures of the previous section offer some clues for how future work could proceed, as well as where listener-error-based and speaker-choice-based accounts might make additional predictions about either ongoing or future perceptibility effects, or about articulatory variation by speakers.

Recall that listener-error based accounts posit that listener-learners make perceptual errors in what they hear and/or misattribute what they hear to phonological rather than phonetic causes. Figure 2.7 indicates which segments are estimated to be out-of-context confusable. Figure 2.11 offers a single, systematic window on which segment types in American English are estimated to have almost no sensitivity in confusability to local phonotactic context, and which, in contrast seem to have higher variability (and therefore sensitivity) across local phonotactic contexts. Those segments that are both typically out-of-context confusable and which are sensitive to phonotactic context seem (all else being equal) like the best candidates to examine for perceptibility effects driven by relatively higher out-of-context confusability, i.e. of the kind described by listener-error accounts. Figure 2.6, however, shows that many of these segment types are not typically that confusable in their actual incremental contexts (with Fig. 2.12 indicating this is usually a result of incremental context), and suggests that only the subset of them that are typically contextually confusable are good candidates to examine in more detail for ongoing or future perceptibility effects of the kind expected under listener-error accounts. Figures 2.9 and 2.10 further suggest that the left edge of the word (and/or segment tokens in shorter words) are the most likely places to examine for such candidates.

Turning to speaker-choice accounts of perceptibility effects, recall that they argue that, on top of what listener-error accounts predict, it is also the case that speakers vary their pronunciations on the basis of their communicative goals, selectively enhancing acoustic features of words that support those goals and reducing others where such reduction does not impact those goals. While the results reported here do not model communicative value (or lack thereof), an intuitively necessary condition for enhancement of a segment token is that it be worth enhancing: both valuable to the goal of communication and counterfactually at risk of impeding or not being as useful to the speaker's communicative goals if left unenhanced. Figures from the previous section indicate that there are some segments or

segment-context combinations that are not at all typically contextually confusable (and hence according a speaker-choice account are both — *ceteris paribus* — least likely to be enhanced, and uniquely good candidates for reduction), and other segments or segment-context combinations that *are* contextually confusable (and hence plausible candidates for enhancement).

2.6.3 Future work

The clearest avenue for follow-up work is development of a more accurate triphone-to-triphone channel model for investigation of the predictions reported in the previous subsection and their relation to reduction and enhancement of particular classes of segments in particular classes of contexts.

Another direction for future work is cross-linguistic investigation — Dutch is the only other language (at present) where comparable confusability data is available (Smits et al., 2003). Given the overall relative typological similarity of Dutch and English, differences between the two languages in the relative perceptibility of their common segments are plausibly attributable to differential top-down effects caused by differences in the structure of their segmental inventories and lexicons.

Figures 2.2a-b are Figures 1-2 from Tanenhaus et al. (1995), and are reprinted with permission from AAAS.

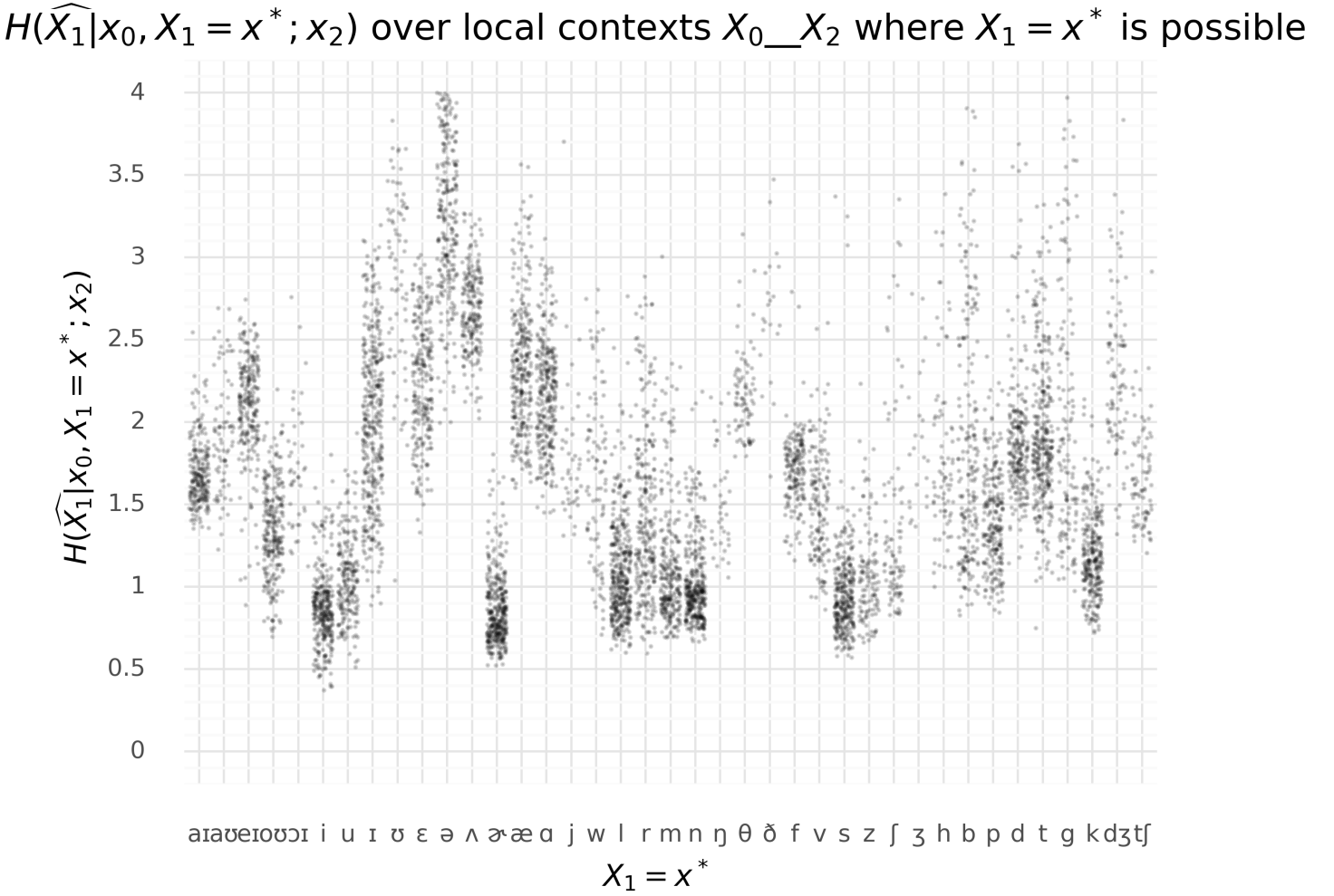


Figure 2.7: Posterior contextual surprisal of each segment type x^* in the artificial lexicon.

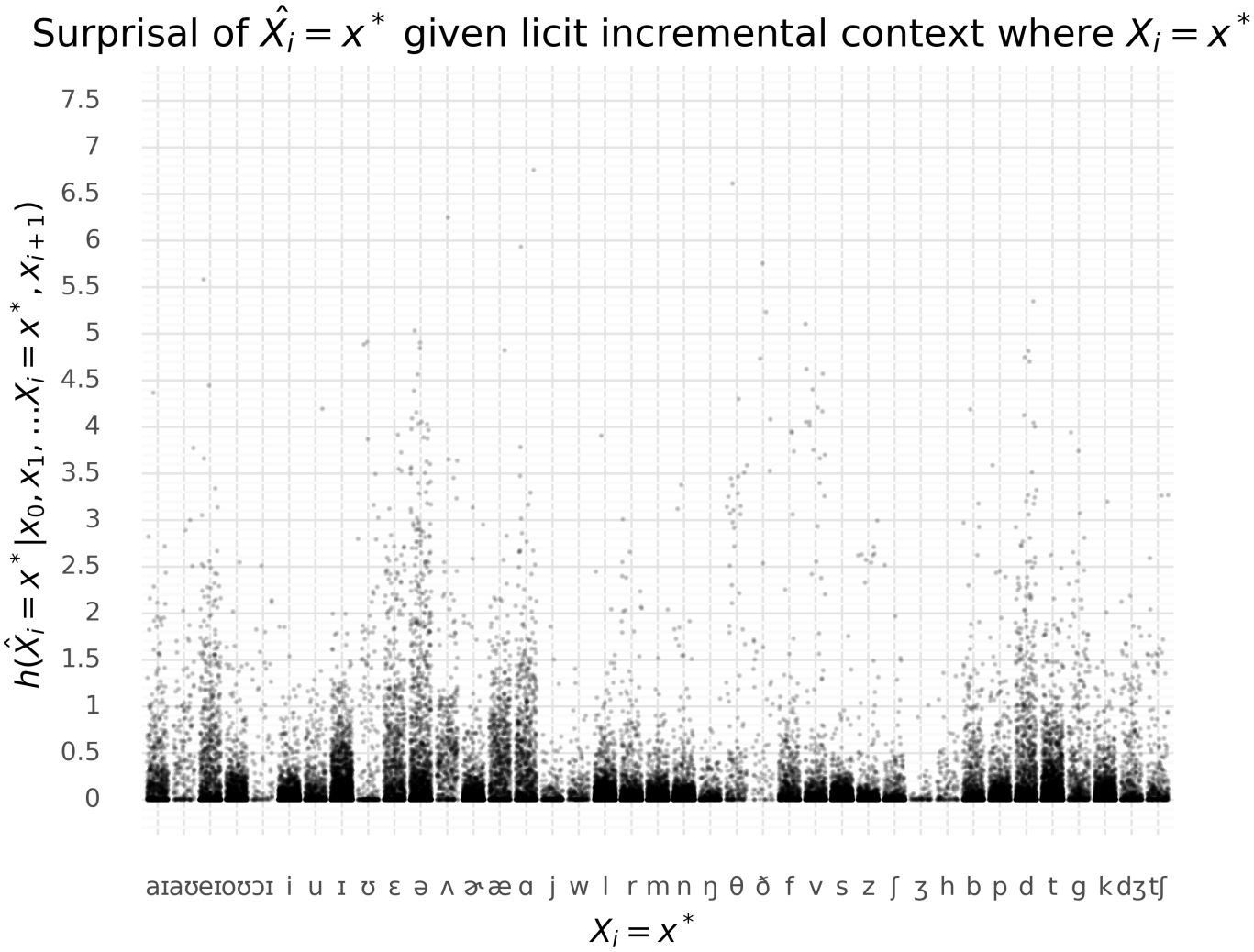


Figure 2.8: Posterior contextual surprisal of each segment type x^* in the natural lexicon.

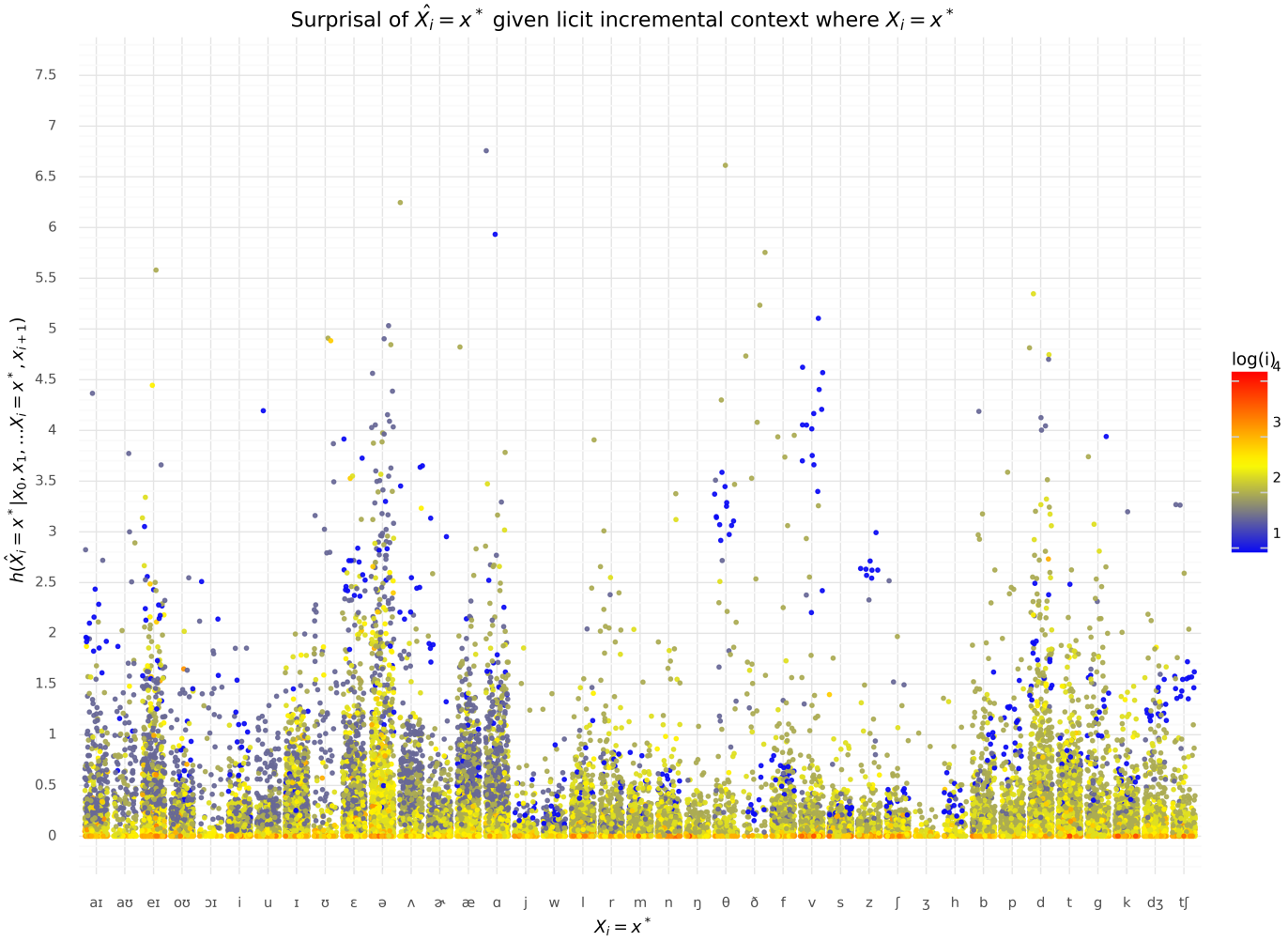


Figure 2.9: Posterior contextual surprisal of each segment type x^* in the natural lexicon, with color reflecting $\log_2$ position within the word (distance from the left edge).

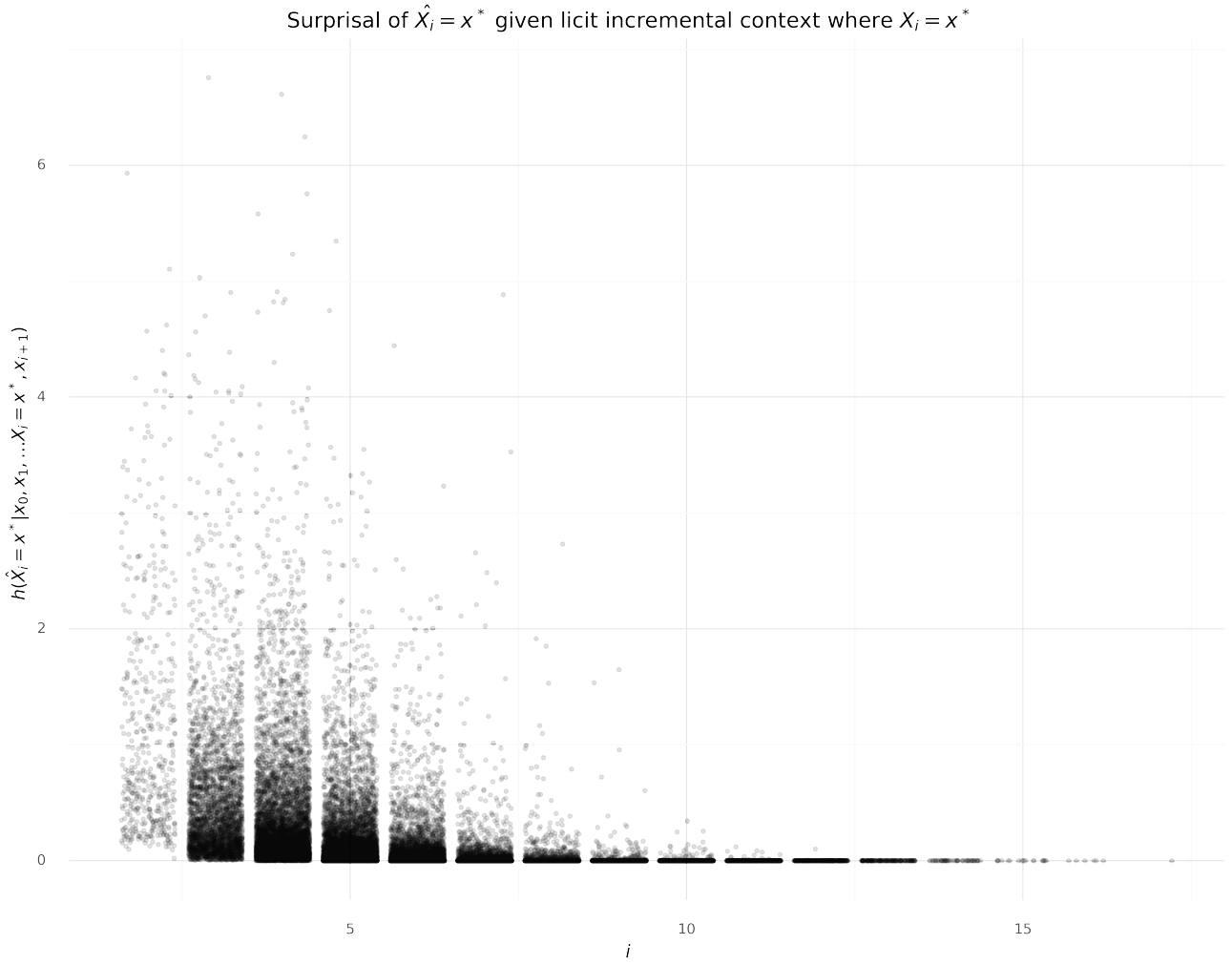


Figure 2.10: Posterior contextual surprisal of individual segment tokens in the natural lexicon as a function of position within the word (distance from the left edge), aggregated over all segment types.

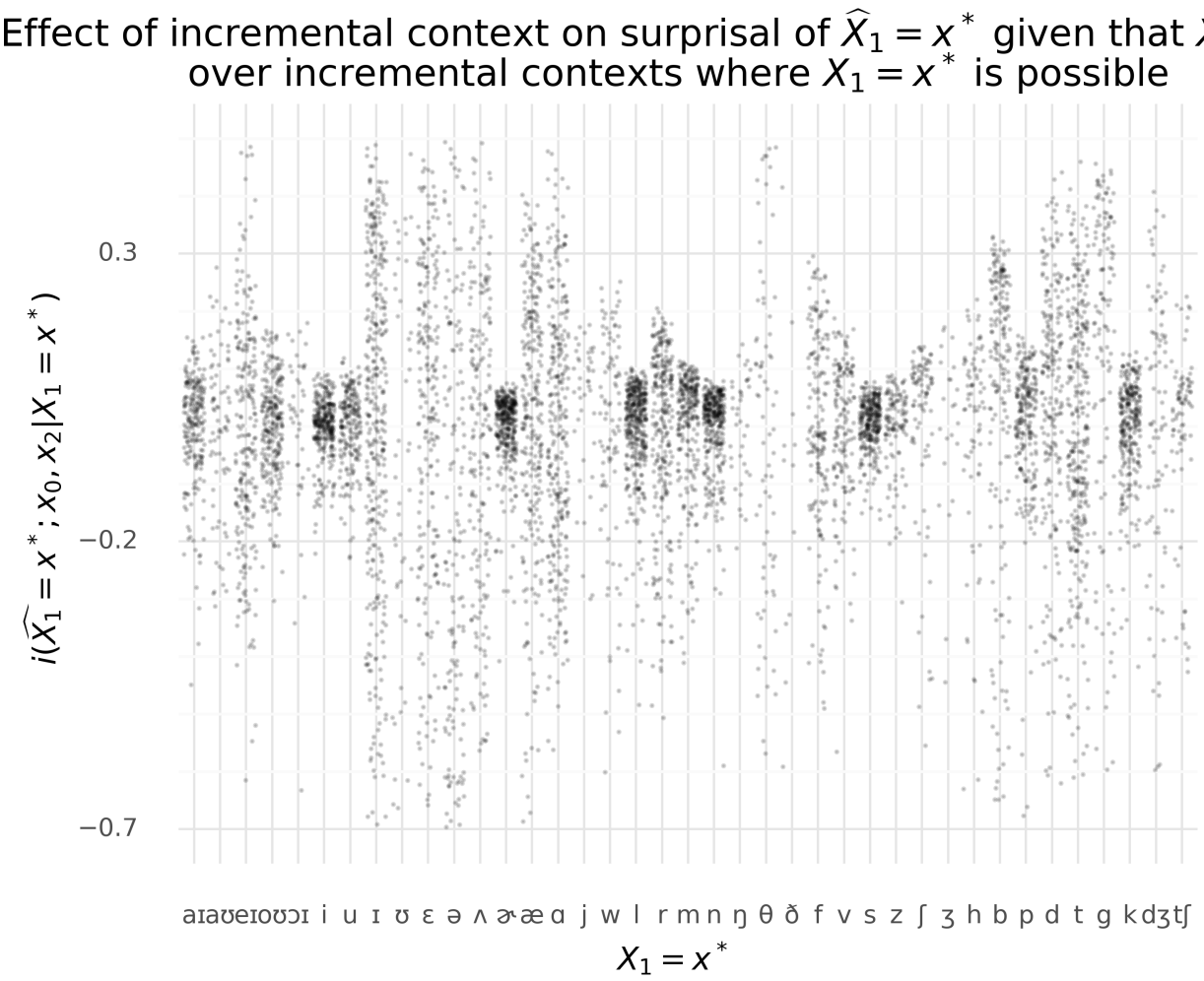


Figure 2.11: Effect of incremental context on perceptibility of each segment type x^* in the artificial lexicon, showing fine detail but omitting some datapoints.

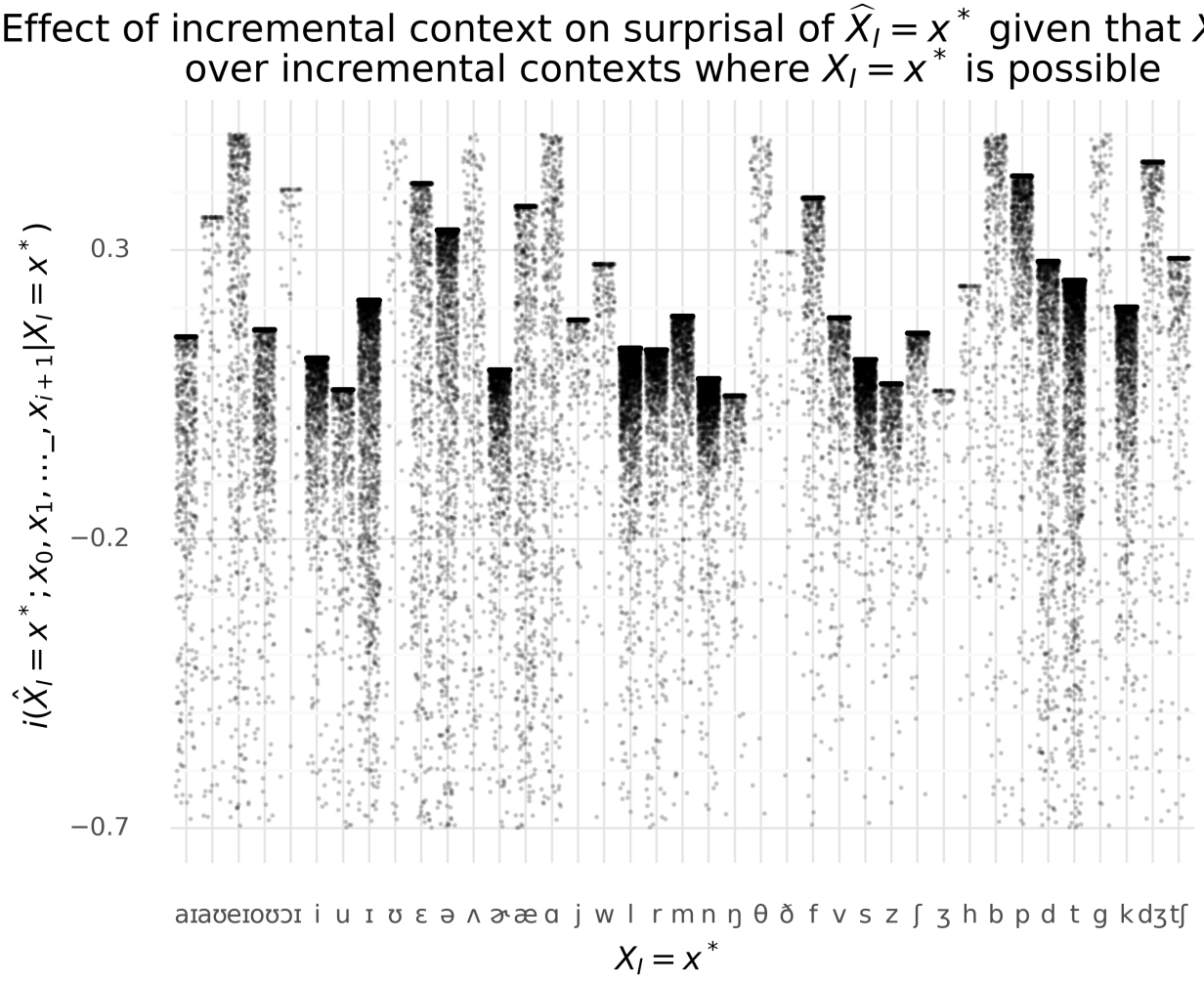


Figure 2.12: Effect of incremental context on perceptibility of each segment type x^* in the natural lexicon, showing fine detail but omitting some datapoints.

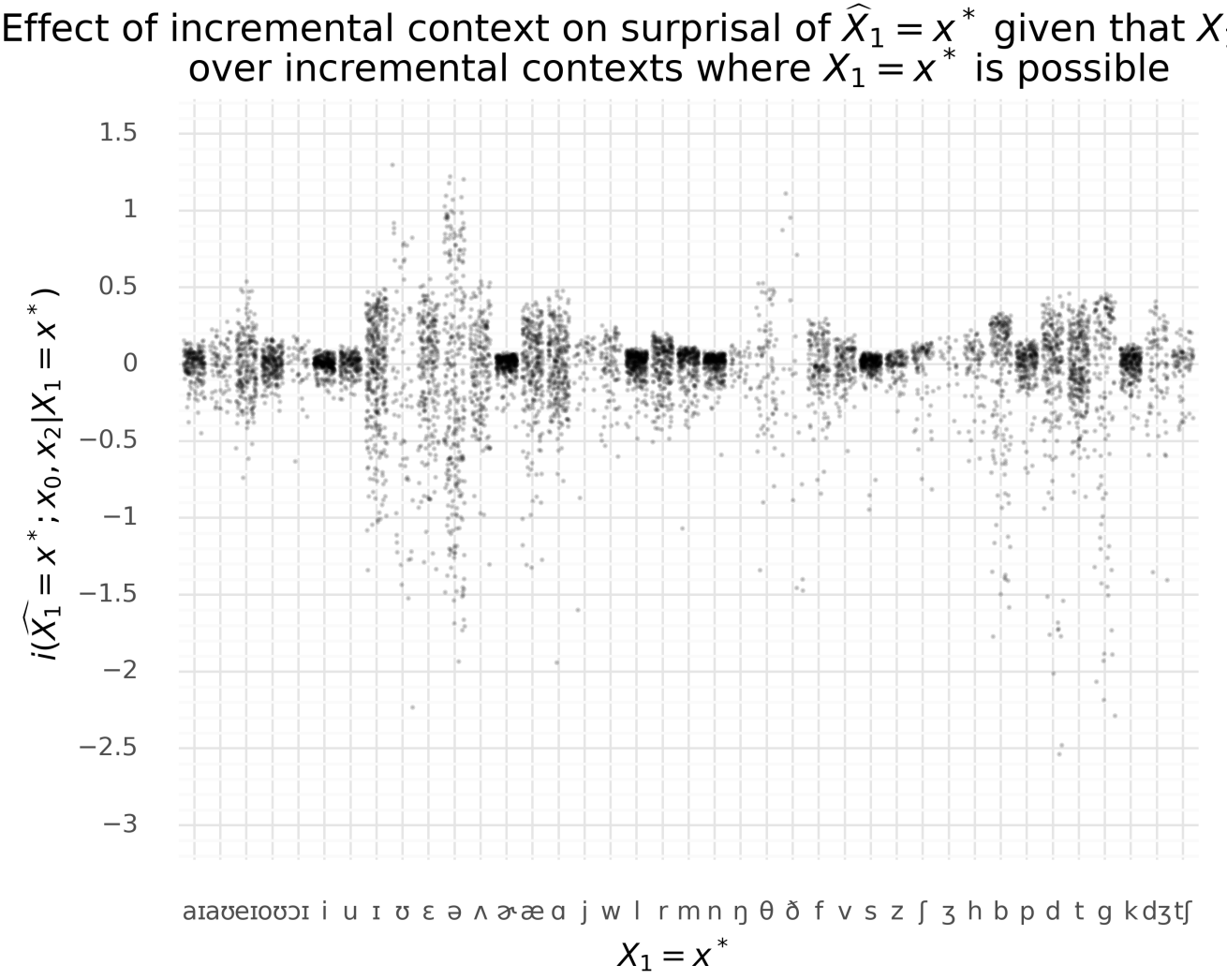


Figure 2.13: Effect of incremental context on perceptibility of each segment type x^* in the artificial lexicon.

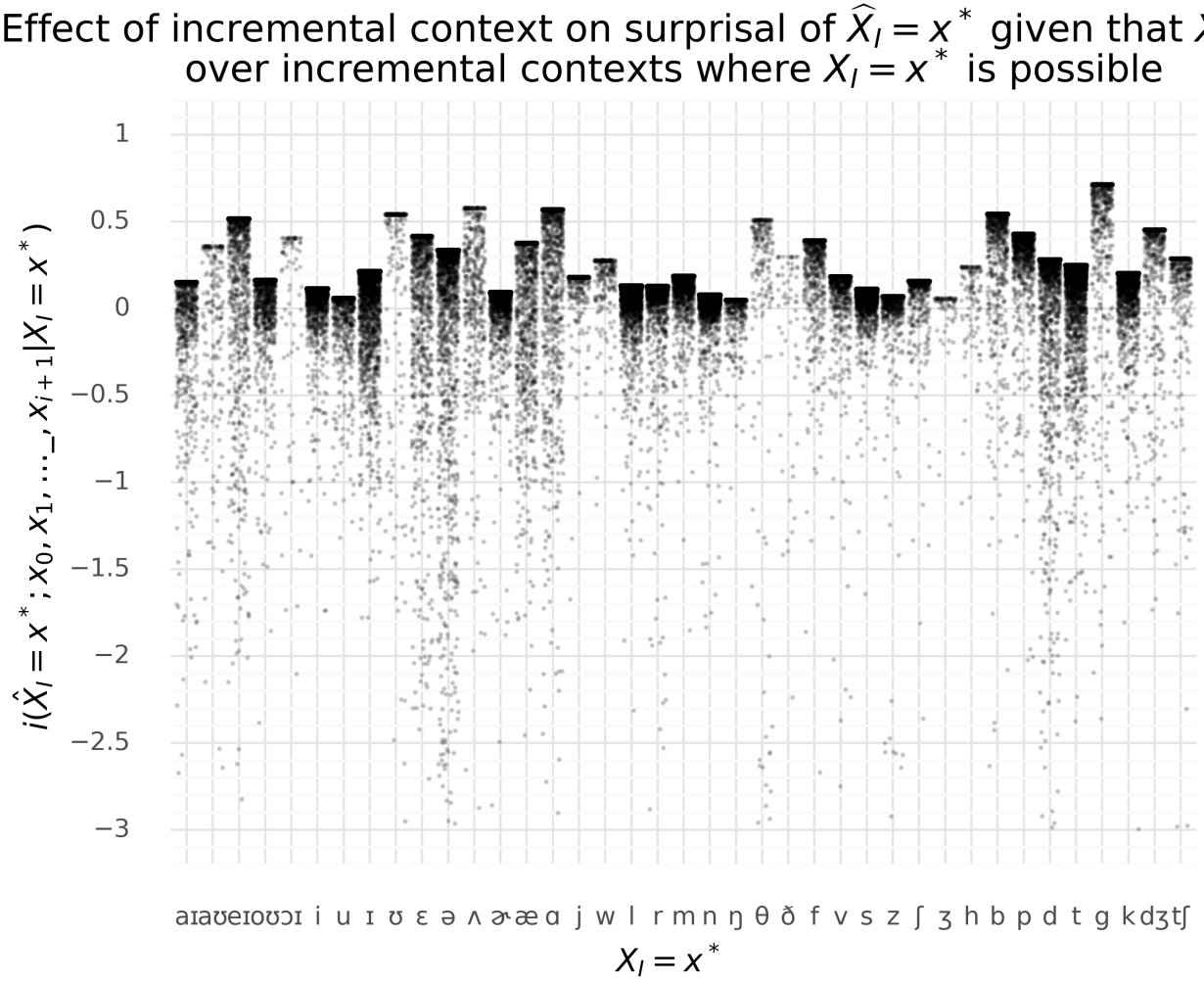


Figure 2.14: Effect of incremental context on perceptibility of each segment type x^* in the natural lexicon, zoomed in and still omitting some datapoints.

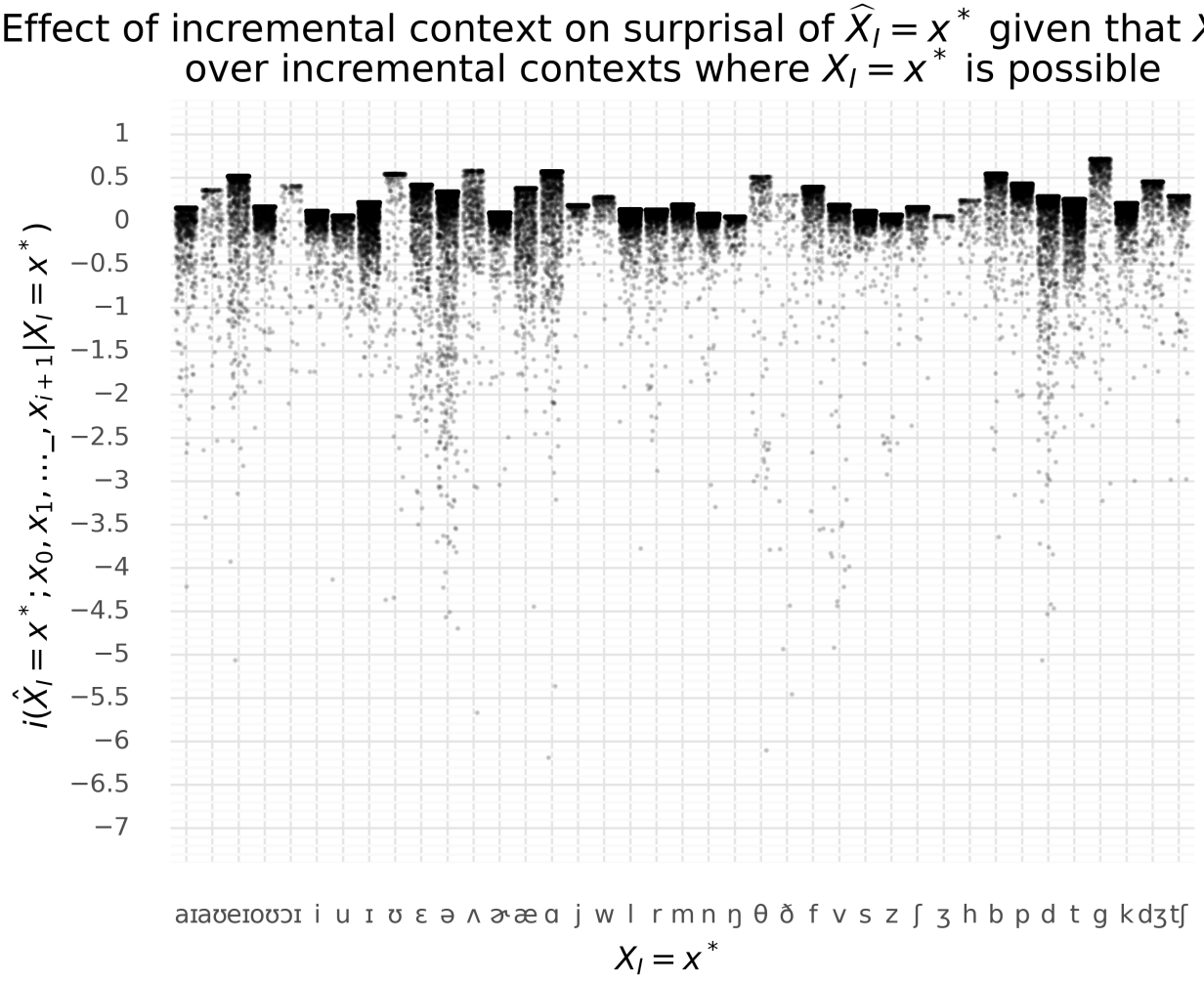


Figure 2.15: Effect of incremental context on perceptibility of each segment type x^* in the natural lexicon, showing the full range of variation.

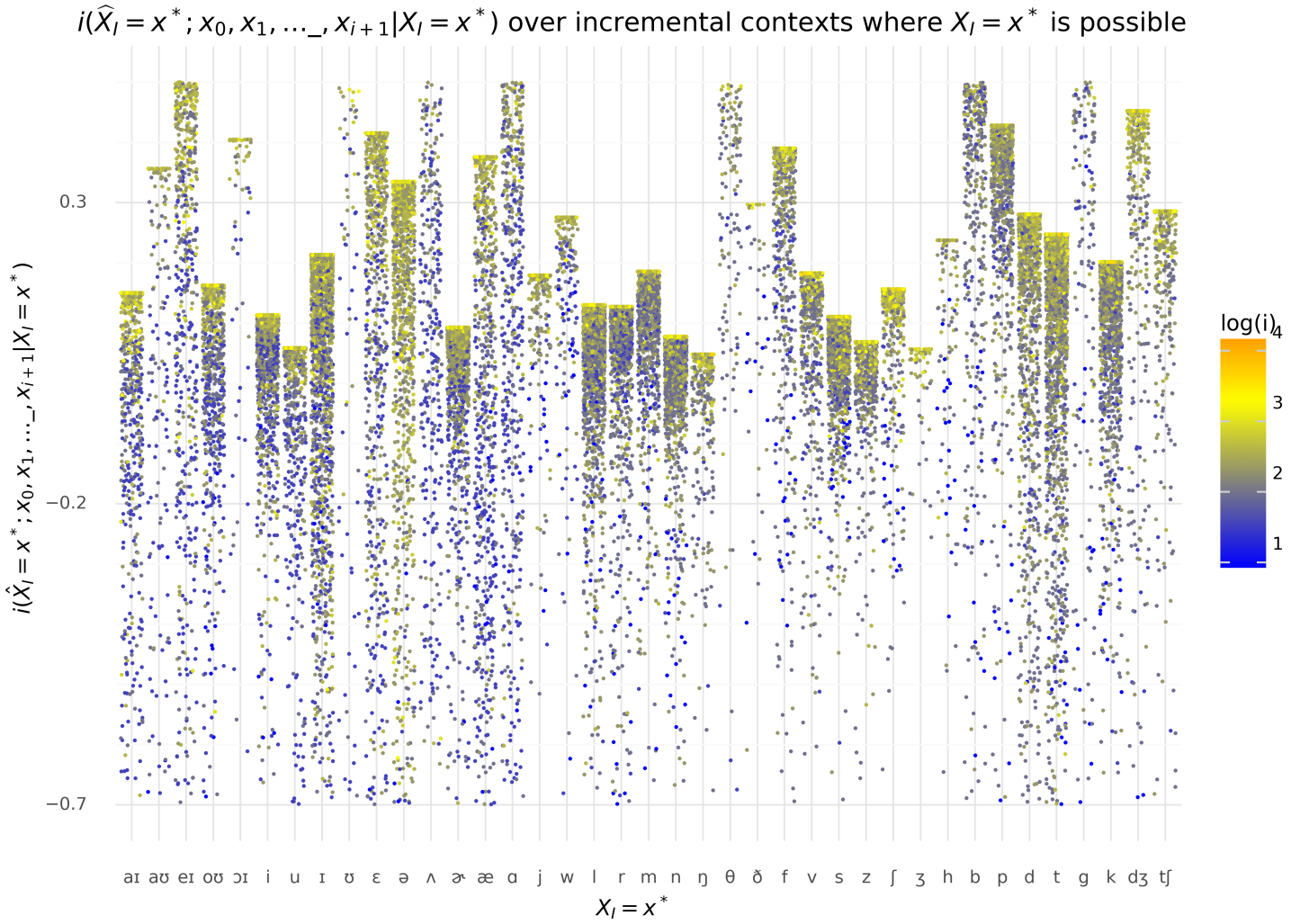


Figure 2.16: Effect of incremental context on perceptibility of each segment type x^* in the natural lexicon, with color reflecting $\log_2$ position within the word (distance from the left edge).

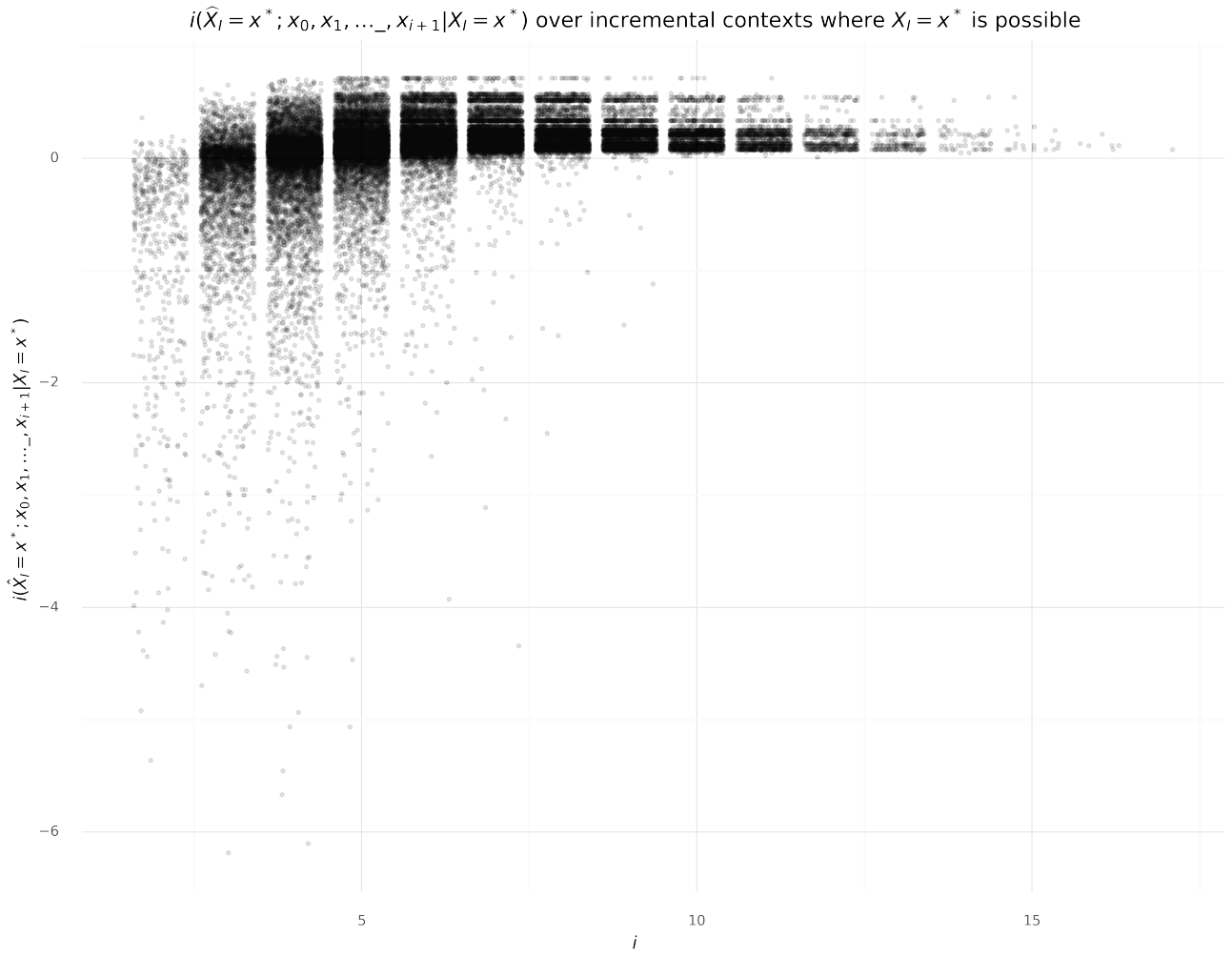


Figure 2.17: Effect of incremental context on perceptibility of segment tokens in the natural lexicon as a function of position within the word (distance from the left edge).

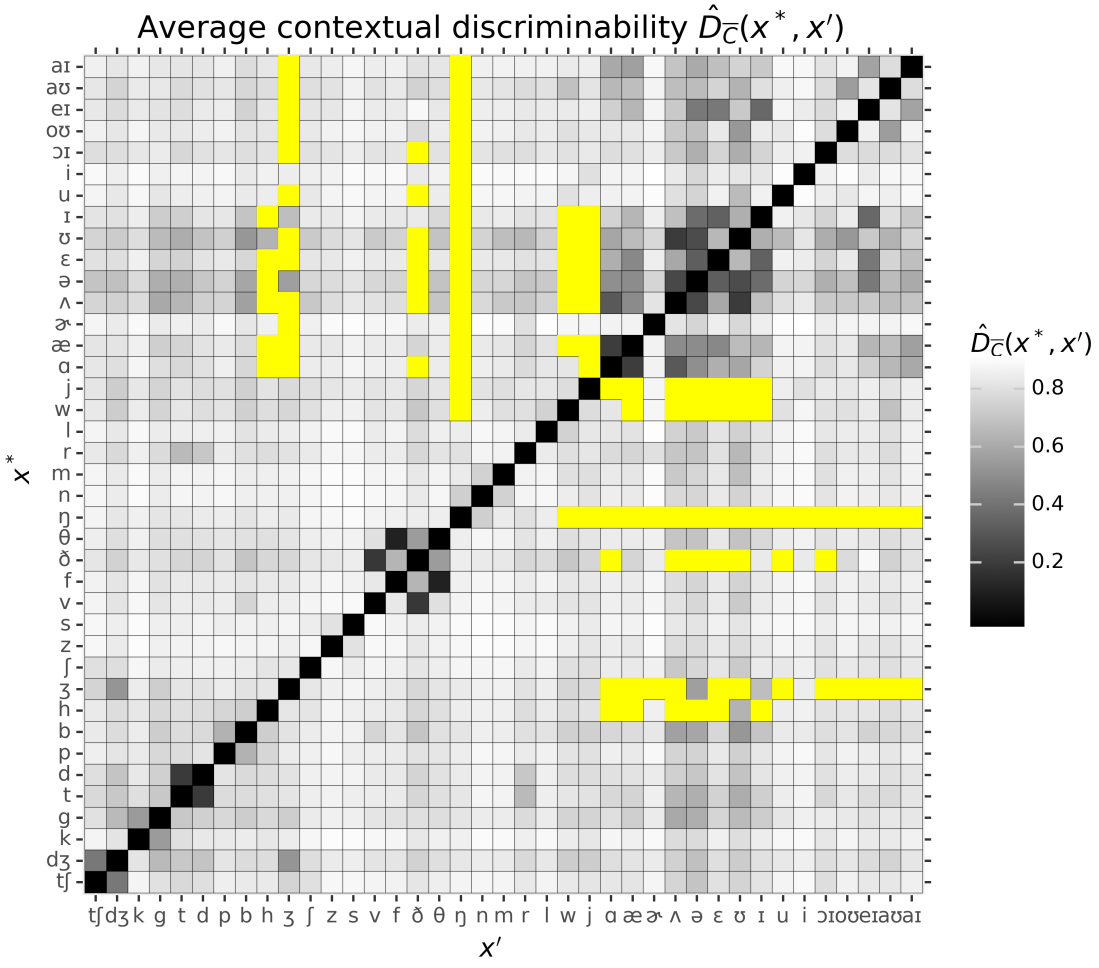


Figure 2.18: Similarity of average contextual confusability in the artificial lexicon for each pair of segment types x^*, x'. Yellow regions indicate no data.

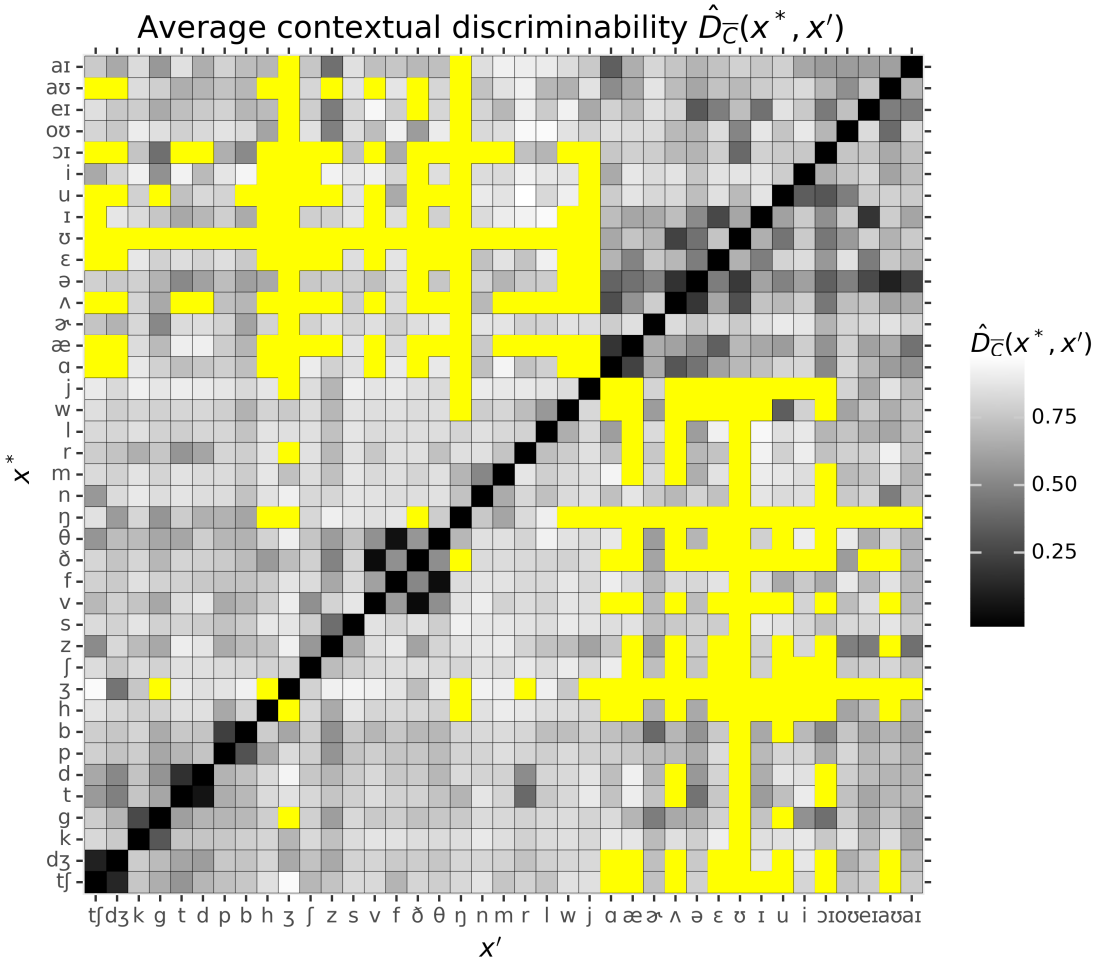


Figure 2.19: Similarity of average contextual confusability in the natural lexicon for each pair of segment types x^*, x'. Yellow regions indicate no data.

Chapter 3

Speakers enhance contextually confusable words

3.1 Introduction

A major open question in the study of natural languages is the extent to which pressures for efficient communication shape the online production choices of speakers or the ('offline') system of forms and form-meaning mappings. Zipf (1936, 1949) famously noted that highly frequent words tend to be shorter and hypothesized that this could be explained in terms of pressures for efficient communication: the average cost of producing a word is lower than it would be otherwise. More recent work has formalized hypotheses about the effect of communicative pressures on language usage and design using tools from information theory (Cover & Thomas, 2012; Shannon, 1948) and rational analysis (Anderson, 1990, 1991). This work has found evidence that meanings are allocated to word types in a way that minimizes speaker effort (Piantadosi, Tily, & Gibson, 2011; Piantadosi et al., 2012), and that this appears to be at least partly explainable by online production choices (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013).

While this research offers evidence that lexicons and the production choices of speakers are shaped by pressures for efficient communication, other work examining how much words and lexicons are shaped by pressures for ensuring effective communication in the face of noise and uncertainty has been more equivocal. For example, pressures for robustness to noise would be expected to cause the words of natural lexicons to be dispersed and distinct from each other, preventing confusions between different words. Dautriche, Mahowald, Gibson, Christophe, and Piantadosi (2017), however, finds that lexicons exhibit clear tendencies towards being clumpier rather than dispersed.

Many studies have used the phenomena of *reduction* and *enhancement* to investigate whether communication is optimized for robustness to noise. Speech tokens that are produced with shorter than usual duration, or with parts omitted or made less distinctive, are said to be reduced, and those tokens produced with longer durations or produced more distinctively are enhanced.

One line of work has provided evidence that contextual predictability influences reduction and enhancement: words, syllables, and segments that are more contextually predictable tend to be reduced and those that are less contextually predictable tend to be enhanced (see e.g. Aylett and Turk 2004, 2006; Buz, Tanenhaus, and Jaeger 2016b; Cohen Priva 2008, 2012, 2015; Demberg, Sayeed, Gorinski, and Engonopoulos 2012; Jurafsky, Bell, Gregory, and Raymond 2001; Pate and Goldwater 2015; Seyfarth 2014; Turnbull, Seyfarth, Hume, and Jaeger 2018; Van Son, Koopmans-van Beinum, and Pols 1998; Van Son and Pols 2003; see Bell et al. 2009; Jaeger and Buz 2018 for reviews). According to a communicatively-oriented account, this is explainable as balancing efficiency against effectiveness: speakers economize on production cost the more that context facilitates accurate listener inference of the speaker's intent.

A second line of work, more closely related to the current study, has examined the effect of *neighborhood density* on reduction and enhancement. This work has found

evidence that words with greater *neighborhood size* or *density* — that is, words that have a greater number of similar-sounding neighbors — have faster onset of production, and have lower overall durations, i.e. they are reduced. Words with greater neighborhood density also take longer for listeners to recognize and comprehend, and have less acoustically distinctive vowels (Gahl et al. 2012; Vitevitch 2002; see Vitevitch and Luce 2016 for review). While neighborhood density has been found to predict a number of behavioral measures, its interpretation and what ultimately drives related effects remains unclear (Gahl & Strand, 2016; Sadat, Martin, Costa, & Alario, 2014; Vitevitch & Luce, 2016).

This second line of work provides a challenge for communicatively-oriented models of production: words with greater numbers of similar-sounding neighbors (or greater average acoustic similarity) seem likely to be more confusable, and therefore speakers would be predicted to decrease the likelihood of noise by, e.g., increasing their duration. However, this work does not directly estimate word confusability, instead using neighborhood density or an acoustic similarity measure as a proxy. It remains possible that greater word confusability is associated with phonetic enhancement, and that a more direct measure of confusability would reveal this relationship.

In this paper, we present the first comprehensive measure of relative word confusability based on both a language model and psychoacoustic data, and we examine how well it predicts word durations in natural speech corpora. We first present a derivation of a Bayesian model of word recognition (broadly similar to Norris and McQueen 2008) that incorporates both linguistic context and a model of noise estimated from the gating data of Warner et al. (2014). We use this speech recognition model to define a measure of confusability, and apply this measure to content words in the NXT-annotated subset of the Switchboard corpus and in the Buckeye corpus (Calhoun et al., 2010; Pitt, Johnson, Hume, Kiesling, & Raymond, 2005). We provide evidence that greater confusability is associated with longer duration.

3.2 A model of word confusability

We propose a simplified model of word confusability, in which there are two factors that will make word v in context c more vs. less confusable. On the one hand, a listener who has observed context c has some 'top-down' beliefs and expectations about what v will be before the speaker produces any acoustics for v. On the other hand, once the speaker has produced acoustics for v, there will be (in general ambiguous) 'bottom-up' acoustic cues that will usually underdetermine what the speaker's choice of v actually was. The goal of the listener is then to combine their top-down expectations with their bottom-up observations to reason about which words are more vs. less likely to have been what the speaker intended.[1]

We operationalize the perceptibility of word v as the probability that the listener accurately recovers this word in situations where the speaker uses it; the confusability of a word is inversely related to its perceptibility. If a speaker has a model of the expected confusability of a given word, they can then decide to lengthen or shorten their particular production of the word token, balancing listener comprehension and their own effort.

3.2.1 Model definition

To model the in-context confusability of word tokens, we model the task of word recognition as one of Bayesian causal inference, with the following underlying generative process for the speaker:

1. At some point in time, the speaker has already produced some existing sentential context c, consisting of a sequence of orthographic words. We assume for simplicity and tractability that the listener knows exactly what this context is at each timestep.

[1]Note that of the two basic factors integrated here, previous probabilistic work on reduction has been limited to using only 'top-down' expectations.

2. The speaker produces the current word v — e.g. *cigarette*. We model this as sampling according to a language model p_L: $v \sim p_L(\cdot|c)$.
3. The speaker determines the segment sequence $x_1^f = (x_1, ..., x_f)$ corresponding to their word choice. For example, the speaker will determine that the segments [sɪgəɹɛt] correspond to the word *cigarette*.

   In our corpora, there is a unique correct segment sequence for a given orthographic word. For ease of exposition, we therefore identify x_1^f with its corresponding orthographic form v. Abusing notation, we will write $p_L(x_1^f|c)$ for the distribution over segmental forms induced by the language model.[2]
4. The listener receives a segment sequence $y_1^f = (y_1, ..., y_f)$ — e.g. [ʃɪgəɹɛt] (‘*shigarette*’) — drawn from a channel distribution p_N conditioned on the speaker’s intended segment sequence: $y_1^f \sim p_N(\cdot|x_1^f)$. This represents the effects of noise on the signal received by the listener.

The task of the listener is to then combine their observation (represented here by y_1^f) with their prior expectations about which words are likely given the context. The listener tries to determine how likely each wordform in the lexicon is to have been the one intended by the speaker. Their posterior belief p_{LISTENER} about which segmental wordform x_1^f was intended is described by Bayes’ rule:

$$p_{\text{LISTENER}}(x_1^f|y_1^f, c) = \frac{p_N(y_1^f|x_1^f)p_L(x_1^f|c)}{p(y_1^f|c)} \tag{3.1}$$

$$= \frac{p_N(y_1^f|x_1^f)p_L(x_1^f|c)}{\sum_{x_1'^f} p_N(y_1^f|x_1'^f)p_L(x_1'^f|c)} \tag{3.2}$$

Suppose for example that the listener perceives y_1^f =[ʃɪgəɹɛt]. Their beliefs about the lexicon

[2]This notation ignores homophony, though the model is in fact sensitive to this.

$p_L(X_1^f|C)$ will tell them that this is not a valid segmental wordform, but that [sɪgəɹɛt] *is* a valid wordform. Their beliefs about the noise distribution for the language $p_N(Y_1^f|X_1^f)$ tell them that x_j =[s] is a plausible segment to be misperceived as y_j =[ʃ]; together this suggests that a good explanation of their percept is the intended wordform x_1^f =[sɪgəɹɛt].

Equation 3.1 allows us to measure how accurately the listener will be able to reconstruct the speaker's intended message, given a perceived segmental wordform y_1^f. However, this is not sufficient to determine the confusability of an intended wordform. In general, an intended wordform x_1^f may give rise to many different perceived wordforms y_1^f as a result of noise. In order to measure its confusability, we therefore need to marginalize over the possible perceived segment sequences.

We define the contextual perceptibility of a segmental wordform x_1^f in context c to be the expected probability that the listener accurately recovers it:

$$\mathop{\mathbb{E}}_{y_1^f \sim p_N(\cdot|x_1^f)} p_{\text{LISTENER}}(x_1^f|y_1^f, c) \tag{3.3}$$

$$= \sum_{y_1^f} p_{\text{LISTENER}}(x_1^f|y_1^f, c) p_N(y_1^f|x_1^f) \tag{3.4}$$

The space of all possible channel strings y_1^f grows exponentially in sequence length f. However, each segment is only substantially confusable with a small number of other segments and the probability of more than a small number of channel errors is small. We therefore approximated Eq. 3.3 with a Monte Carlo estimator:

$$\mathop{\mathbb{E}}_{y_1^f \sim p_N(\cdot|x_1^f)} p_{\text{LISTENER}}(x_1^f|y_1^f, c) \approx \frac{1}{n} \sum_{i=1}^{n} p_{\text{LISTENER}}(x_1^f|y_{1,i}^f, c) \tag{3.5}$$

$$y_{1,i}^f \sim p_N(\cdot|x_1^f) \tag{3.6}$$

We choose $n = 1000$ to balance the variance and computational feasibility of the estimator.

Finally, following the reasoning given in Levy (2005, 2008a), we take the negative logarithm of this quantity and arrive at a surprisal, which represents the contextual confusability of segment sequence x_1^f in context c:[3]

$$h(x_1^f | x_1^f, c) = -\log \mathop{\mathbb{E}}_{y_1^f \sim p_N(\cdot|x_1^f)} p_{\text{LISTENER}}(x_1^f | y_1^f, c) \tag{3.7}$$

3.3 Materials and methods

We make use of two types of data: psychoacoustic gating data for estimating a noise model, and several corpora of natural speech for evaluating whether individuals increase the duration of more confusable words.

3.3.1 Words duration data

Word durations were analyzed separately in two spoken corpora of American English: the Buckeye Corpus of Conversational Speech (Pitt et al., 2005) and the NXT Switchboard Annotations (Calhoun et al., 2010), a highly annotated subset of Switchboard-1 Release 2 (Godfrey & Holliman, 1997).

The Buckeye Corpus contains about 300,000 word tokens, taken from interviews with 40 speakers from central Ohio. Word durations for the present study were taken from the timestamps provided for word-level annotations. Each word token had a broad transcription uniform across all instances of the word type and a second, token-specific close transcription created by a human annotator.

The Switchboard Corpus contains transcripts of telephone conversations between strangers. The NXT annotated subset includes about 830,000 word tokens from 642 conver-

[3]Compare Equations 3.3–3.7 with Eq. VII of Levy (2008b), a study of sentence-level confusability.

sations between 358 speakers recruited from all areas of the United States. Word durations for the present study were taken from the 'phonological word'-level timestamps; these were the result of annotator-checked and -corrected timestamps initially made by alignment software. Each phonological word was also associated with a segmental transcription that was uniform across all instances of the word type.

Exclusion criteria almost exactly follow Seyfarth (2014) for the reasons cited there. These criteria are mainly designed to exclude non-content words and words whose pronunciation is likely affected by disfluencies or prosodic structure. Our criteria only diverge in the following manner: Word tokens were excluded if the utterance speech rate (total number of syllables / length of the utterance in seconds) was more than 3 standard deviations from the speaker mean (vs. 2.5 in Seyfarth 2014). After exclusion criteria were applied, about 44,000 (4,900) and 113,000 (8,900) word tokens (word types) remained in the Buckeye and NXT Switchboard corpora, respectively.

3.3.2 Diphone gating data

The model of word confusability was based on the diphone gating experiment data of Warner et al. (2014). Participants listened to gated intervals of every phonotactically licit diphone of (western) American English and attempted to identify the full diphone they thought was being produced during the interval. Along with earlier work by some of the same researchers on Dutch (Smits et al., 2003; Warner et al., 2005), this represents by far the richest and most comprehensive acoustic confusion matrix data of its kind.

Warner et al. (2014) identified all adjacent pairs of segments within and between words based on an electronic pronouncing dictionary of about 20,000 American English wordforms. A set of approximately 2,000 phonotactically licit diphones were extracted from this transcribed lexicon. At least one stimulus nonsense word was created per diphone by inserting the diphone into an environment consisting of at most one syllable on the left

and at most one syllable on the right.

A recording of each stimulus wordform was then marked up with (generally) six temporal gates. For each stimulus wordform, one recording was created for each gate, starting at the beginning of the original recording and going all the way up to a gate location, followed by a ramping procedure (rather than truncation or white noise) to avoid systematically biasing confusion data.

In each trial, participants heard a gated stimulus recording.[4] If the recording included a preceding context, this context was displayed on the screen. The participant then selected the stimulus diphone they thought was in the recording (i.e. not including context).

From this response data, each gate of each stimulus diphone can be associated with a frequency distribution over response diphones. Only the response data for gates corresponding to the end of each segment of the diphone were used in the current study. For each of Buckeye and NXT Switchboard, the segment inventories of the gating data and of each speech corpus had to be projected down to a common set of segments. In each case, this involved collapsing the distinction in the corpora between syllabic and non-syllabic nasal stops. For reasons of data sparsity, the distinction between stressed and unstressed versions of any given vowel was also collapsed.

3.3.3 Language model

Our measure of contextual confusability uses a language model to compute the prior probability of a word in context. We estimate a language model from the Fisher corpus (Cieri, Miller, & Walker, 2004), a speech corpus matched for genre and register to Buckeye and Switchboard. This corpus contains about 12 million (orthographic) word tokens taken from nearly 6000 short conversations, each on one of about 100 topics.

We estimated n-gram models of several orders from the Fisher corpus using

[4] See Grosjean (1980) for reference on the gating paradigm.

KenLM (Heafield, 2011).[5] The *n*-gram order was treated as a hyperparameter, and selected on the Training Set, as described below. An add-1 smoothed unigram model was also created from word frequencies in the Fisher corpus using SRILM (Stolcke, 2002; Stolcke, Zheng, Wang, & Abrash, 2011).

3.3.4 Channel model

The *channel model* describes the conditional distribution $p_N(Y_1^f | X_1^f)$ over what sequence of segments y_1^f a listener will perceive (e.g. [ʃɪgəɹɛt], *shigarette*) given the full intended sequence x_1^f (e.g. [sɪgəɹɛt], *cigarette*). We estimate this distribution using the diphone gating data in Section 3.3.2. We make the simplifying assumption that the channel distribution for segment y_i is conditionally independent of all other y_j ($j \neq i$) given intended segments x_{i-1}, x_i, x_{i+1}.

By conditioning on adjacent segments, we can capture some effects of coarticulation on confusability. For example, nasals before oral stops are systematically likely to be misheard as having the same place of articulation as the stop: x_1^f =[ɑnpɑ] (alveolar nasal before labial stop) is more likely to be misperceived as y_1^f =[ɑmpɑ] (a labial nasal) than the reverse, and a confusion of [n] for [m] is comparatively less likely when [n] is between vowels as in [ɑnɑ] (J. J. Ohala, 1990a).

For each gate $g \in \{3, 6\}$ and for each diphone x_1x_2, the response data from Section 3.3.2 induce a conditional frequency distribution over channel diphones $f_g(y_1, y_2|x_1, x_2)$. These frequency distributions were smoothed by adding a pseudocount to every channel diphone in every distribution; the distributions were then normalized to define a smoothed pair of diphone-to-diphone channel distributions $p_g(y_1, y_2|x_1, x_2)$. From the marginals of these distributions we constructed an approximation (Eq. 3.8) of the triphone-to-uniphone

[5] We do not use lower-perplexity neural language models due to intractability resulting from the normalizing constant in Equations 3.2 and 3.3.

channel distribution via their geometric mean:[6]

$$\widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1}) \propto \sqrt{p_3(y_i|x_{i-1}, x_i) \cdot p_6(y_i|x_i, x_{i+1})} \tag{3.8}$$

With the simplifying assumption that only substitution errors are possible,[7] we obtain a preliminary string-to-string channel model:

$$\widetilde{p}_N(y_1^f|x_1^f) = \prod_{j=1}^{j=f} \widetilde{p}_t(y_j|x_{j-1}, x_j, x_{j+1}) \tag{3.9}$$

We are primarily interested in using the channel model to define a ranking on the confusability of words, i.e. to determine which words are more or less confusable than others. This makes the channel model defined by Equations 3.8 and 3.9 not fully adequate.

The diphone gating data were collected in a laboratory setting with rates of noise lower than for naturalistic speech. As a result, when the noise model is estimated from this data, it implies the absolute rate of accurate perception (as defined by Equation 3.2) is close to 1 for most words. This makes it hard for the Monte Carlo estimator defined in Equation 3.5 to determine stable rankings of confusability. In order to estimate rankings in a more stable manner, we introduce a model hyperparameter $0 < \lambda \leq 1$, and define a new triphone-to-uniphone channel distribution by:

$$p_t(y_i|x_{i-1}, x_i, x_{i+1}) = \left\{ \begin{array}{ll} \lambda \cdot \widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1}), & y_i = x_i \\ \beta \cdot \widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1}), & y_i \neq x_i \end{array} \right\} \tag{3.10}$$

Here $\beta \geq 1$ is used to normalize the distributions; it is fully determined by λ for a particular distribution $p_t(\cdot|x_{i-1}, x_i, x_{i+1})$. The term λ is used to increase the noise rate in the channel

[6] We stop short of utilizing a full triphone-to-triphone channel distribution for tractability.

[7] The gating data does not provide information for estimating the probability of deletion or insertion errors.

distributions. Note that two important features of the original triphone-to-uniphone distributions $\widetilde{p}_t$ are maintained in the new model. First, the ratios of outcome probabilities within a single triphone distribution remain the same:

$$\frac{p_t(y_i|x_{i-1}, x_i, x_{i+1})}{p_t(y'_i|x_{i-1}, x_i, x_{i+1})} = \frac{\widetilde{p}_t(y_i|x_{i-1}, x_i, x_{i+1})}{\widetilde{p}_t(y'_i|x_{i-1}, x_i, x_{i+1})} \tag{3.11}$$

for segments $y_i, y'_i \neq x_i$. Second, the relative probability of accurate perception is preserved across triphone distributions:

$$\frac{p_t(x_i|x_{i-1}, x_i, x_{i+1})}{p_t(x'_i|x'_{i-1}, x'_i, x'_{i+1})} = \frac{\widetilde{p}_t(x_i|x_{i-1}, x_i, x_{i+1})}{\widetilde{p}_t(x'_i|x'_{i-1}, x'_i, x'_{i+1})} \tag{3.12}$$

The new model therefore preserves information about the relative noisiness of different segments, and about which segments are most likely to be confused for each other.

The final string-to-string channel model is defined by:

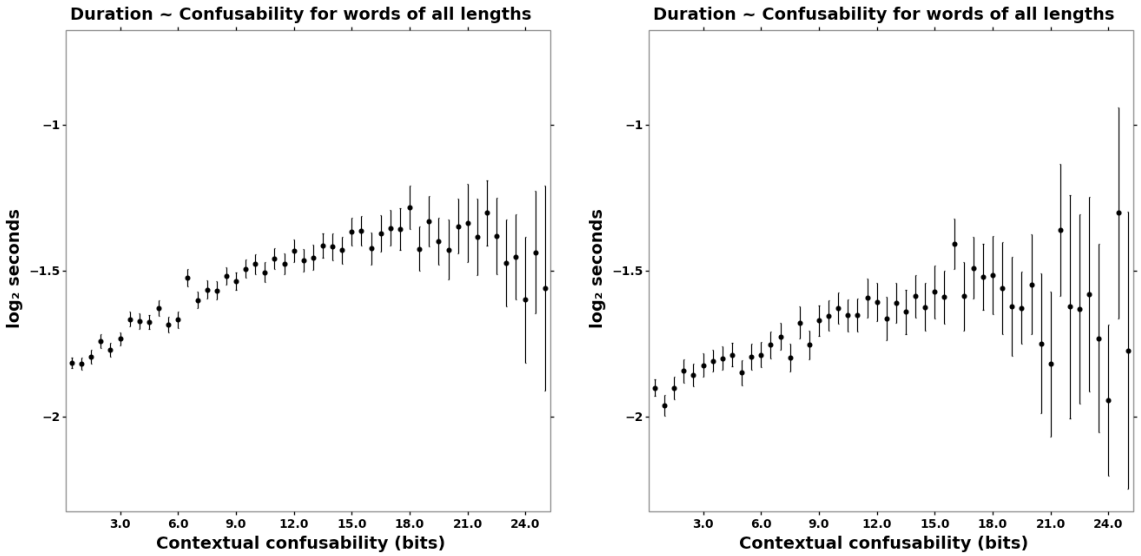
$$p_N(y_1^f|x_1^f) = \prod_{j=1}^{j=f} p_t(y_j|x_{j-1}, x_j, x_{j+1}) \tag{3.13}$$

This new channel model has an increased noise rate, making it easier to estimate stable rankings of confusability across words.

The most similar previous channel model (Norris & McQueen, 2008) was based on Dutch gating data (Smits et al., 2003) comparable to that used here. Norris and McQueen (2008) did not construct a triphone-to-uniphone channel model, but made use of all gates and also allowed investigation of word boundary identification.

3.3.5 Statistical methods

Prior to any analyses, the Switchboard and Buckeye corpora were each randomly divided into evenly-sized Training and Test sets. The Training sets were used for exploratory



(a) Switchboard (b) Buckeye

Figure 3.1: Confusability vs. log duration on the Test sets of the Switchboard and Buckeye corpora. Error bars are 95% confidence intervals (non-bootstrapped). As illustrated in Figure 3.2, data are sparse beyond 18 bits, resulting in large confidence intervals in this range.

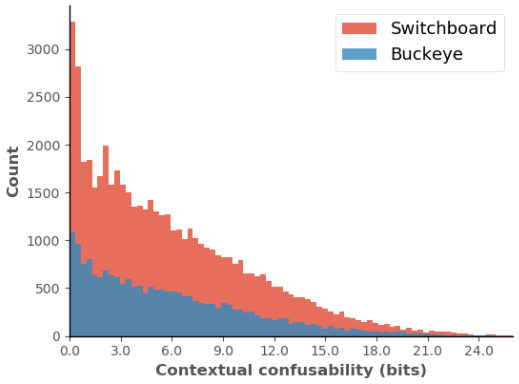


Figure 3.2: Histogram of contextual confusability scores on the Test sets.

statistical analyses, and for determining the values of several model hyperparameters. Following this, all parameters and statistical analyses were frozen, and preregistered with the

Open Science Foundation.

We perform several linear regressions in order to determine the effect of confusability on word duration. Contextual confusability is defined throughout using Equation 3.7. Word durations are log-transformed. The following covariates are standard in the literature, and are included in our analyses: speaker identity; part of speech; unigram prior surprisal; speech rate (the average rate of speech, in syllables per second, of the utterance containing the target word); word length (measured by number of segments and syllables). Several covariates that are included are more non-trivial, and are discussed in more detail below: segmental inventory factors; forward and backward surprisal; neighborhood size and log weighted neighborhood density; and unigram confusability.

The segmental inventory variables code each word as a 'bag-of-segments.' A separate variable is defined for each phoneme in the segmental lexicon of the corpus. Each variable counts the number of times the corresponding phoneme occurs in the word. This is a variant of the baseline model used in previous work (Bell et al., 2009; Gahl et al., 2012).

Certain segments take longer to pronounce than others, and the baseline model is used in case the confusability scores contain information about segment identities within a word. Note, however, that this is a conservative baseline, as segment identity has an effect on confusability; certain segments are, individually, harder to perceive than others. The model will be used to predict word durations after these segmental effects have been factored out.

The forward language-model surprisal of a word is the surprisal of the word given preceding words in the context, and its backward surprisal is the surprisal given the following words in the context. Previous work in English has found backward surprisal to be a stronger predictor of spoken word duration than forward surprisal (Bell et al., 2009; Seyfarth, 2014). Word confusability is expected to be correlated with surprisal, as more surprising words will be more difficult for the listener to recover in the presence of noise.

Neighborhood size and log weighted neighborhood density are measures of the number of words adjacent (within Levenshtein distance 1) to a target word. These measures have been extensively studied as explanatory variables for word duration (see Gahl et al. 2012; Vitevitch and Luce 2016 for review), and are expected to correlate with word confusability: words with more neighbors are expected to be more confusable. We evaluate whether there is any residual effect of confusability beyond its impact on these variables.

Unigram confusability measures the confusability of a word (Equation 3.7) given a unigram (word frequency) language model. This is a measure of the out-of-context confusability of a word, as discussed below.

All variables are treated as fixed effects, and OLS is used for regressions. Confidence intervals and p-values are calculated using the bias-corrected bootstrap. Random effects are not used due to potential issues arising in observational studies like the current one. In particular, random effects may correlate with predictors in an observational study, leading to incorrect estimates of uncertainty and the potential for bias (Bafumi & Gelman, 2006; Wooldridge, 2010).

All analyses were performed in two ways: using the raw values for each variable, and with rank-transformed values for the continuous variables. Our hypothesis is that confusability is monotonically related to word duration. Rank-transformed analyses provide a direct test of the hypothesis.

3.4 Results

Four model hyperparameters were selected using the Switchboard and Buckeye Training sets: the order and direction of the n-gram model, the diphone-to-diphone channel pseudocounts, and the noise factor λ.[8] Backward bigram language models were found to perform best on the Training sets, possibly due to distributional differences between these

[8]The language model order was the same across all covariates where it was used.

corpora and the Fisher corpus, which was used for language model estimation. This is consistent with prior work in the area (e.g. Bell et al. 2009; Seyfarth 2014). Pseudocounts were set to 0.01, and the term λ was set to 2^{-6}.

Figure 3.2 shows the frequency of model-computed confusability scores on the Switchboard and Buckeye Test sets. Figure 3.1 shows the relationship between confusability and word duration on the Test sets.

The first set of analyses include all of the covariates from Section 3.3.5, except for unigram confusability. This allows us to determine whether there is an effect of word confusability on duration, independent of whether this effect is sensitive to context. Greater confusability is associated with longer word durations on both the Switchboard and Buckeye Training sets ($p<0.001$ for all analyses). Table 3.1 shows results of the same analyses performed on the Test sets. The effects replicate on the Test sets, and are qualitatively similar when continuous variables are rank-transformed.

These analyses provide evidence that higher confusability is associated with longer word duration. In the second set of analyses, we investigate whether a context-sensitive measure of confusability is necessary for explaining this effect, or whether an out-of-context measure suffices. In order to do this, we include unigram confusability as a covariate in the analyses, in addition to the previous covariates. Unigram confusability is identical to our target measure of word confusability, except that the language model is replaced with a unigram model. The measure calculates a word's confusability based on its acoustic properties and its phonological similarity to other words. It therefore does not take into account top-down expectations based on a word's context.

After controlling for unigram confusability, contextual confusability remains associated with longer word durations on both the Switchboard and Buckeye Training sets ($p<0.001$ for all analyses). Table 3.2 shows the same analyses on the Test sets. The effects replicate on both Test sets, and similarly for the rank-transformed analyses.

Table 3.1: Effect of contextual confusability on log word duration, not controlling for unigram confusability. Estimates from the Test sets. Rank indicates whether continuous variables were rank-transformed. p-values are upper-bounds.

Dataset	Rank	β	95% CI	p-value
SWBD	No	0.006	(0.004, 0.008)	0.001
SWBD	Yes	0.086	(0.067, 0.109)	0.001
Buckeye	No	0.005	(0.001, 0.008)	0.01
Buckeye	Yes	0.123	(0.080, 0.130)	0.001

Table 3.2: Effect of contextual confusability on log word duration, controlling for unigram confusability. Estimates from the Test sets.

Dataset	Rank	β	95% CI	p-value
SWBD	No	0.009	(0.006, 0.011)	0.001
SWBD	Yes	0.132	(0.095, 0.130)	0.001
Buckeye	No	0.007	(0.003, 0.011)	0.001
Buckeye	Yes	0.148	(0.106, 0.164)	0.001

Chapter 3 was coauthored with Eric Baković and Leon Bergen, and is very similar to the submitted manuscript that has since been edited and published in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1991-2002). The dissertation author was the primary investigator and author of this paper.

Chapter 4

Morphology gets more and more complex, unless it doesn't

4.1 Introduction

Consider the following two large questions that have become central to discussions in the morphological literature, and whose answers are foundational for theory construction in this domain:

(4.1.30) a. What do analysts mean when they talk about morphological complexity and make claims about learnability?

b. What kinds of explanations do analysts advance given what they believe to be true about complexity and social conditions?

Recent work in morphology (Ackerman & Malouf, 2013; Stump & Finkel, 2013) has conceptualized morphological complexity in terms of two interdependent dimensions, ENUMERATIVE COMPLEXITY and INTEGRATIVE COMPLEXITY, henceforth abbreviated as 'E-complexity' and 'I-complexity'. The E-complexity of a language identifies

- the types and numbers of morphosyntactic categories — e.g. tense, case, number…
- the number and shape of formatives used to encode them, and
- the combinatorics and classifications (conjugations or declensions) of those formatives as utilized in the language.

The I-complexity of a language, in contrast, measures the (inter)predictability among wordforms — i.e., it reflects the ways that the enumerative ingredients cataloged by E-complexity are organized into systems of relatedness among classes of words. This kind of complexity has been a central concern of word-based implicative and network approaches — see e.g. (Ackerman, Blevins, & Malouf, 2009; Blevins, 2016; Bonami & Beniamine, 2016; Bonami & Henri, 2010; Bonami & Strnadova, 2018; Bybee, 1985; Janda & Tyers, 2018; Sims, 2015; Sims & Parker, 2016; Wurzel, 1987). An hypothesis associated with this division is that a language can, in principle, vary limitlessly in terms of its E-complexity as long as these ingredients are organized in ways that lead to low conditional entropy (LCE; a measure of I-complexity) for the networks of relations between words constitutive of the morphological system.[1] One aspect of this approach is that, while words exhibit internal structure, the nature of that structure is not necessarily morphemic, as typically assumed in familiar generative frameworks. Rather, word structure is defined by discriminability between (classes of) words and the patterns produced by distinctive arrangements of word elements, i.e., segments, suprasegments, that cohere into systems that constitute language particular systems.[2]

The central explanatory value of systemic organization for morphological phenomena and learnability makes modern linguistic analysis a beneficiary of the early in-

[1]Of course, LCE is likely only one, if important and newly explored, dimension guiding morphological organization.

[2]It is important to note that standard morpheme constructs are subsumed under the discriminability view, since the presence of a morpheme obviously counts as one strategy for distinguishing one (class of) word from another. See Ramscar, Dye, Blevins, and Bayyaen (2018) for discussion of discriminative learning.

sights of paradigm-oriented thinkers like Paul, Kruszewski and de Saussure.[3] For example, Kruszewski (1995) viewed the morphological system as facilitating two fundamental aspects of language usage: *reproduction* was the more or less faithful utterance of stored lexical representations, i.e. fully derived and inflected wordforms and their penumbra of variants, while *production* was the utterance of novel wordforms licensed by the analogical inferences intrinsic to networks of related words.

> . . . every word is connected by twofold bonds: by innumerable ties of similarity with its relatives according to sounds, structure, or meaning and by equally numerous ties of contiguity with its various fellow travellers in every possible kind of phrase. A word is always a member of certain nests or systems of words and at the same time is a member of certain series of words. This explains the ease with which we memorize and recall words. Moreover, these properties of words make it possible for us not to have to resort to straight memorization every time. It is sufficient for us to know words like *idu* ["(I) am walking"], *idës'* ["(you sg.) are walking"], or *vedu* ["(I) am leading"] in order to produce the new word *vedet* ["(he) is leading"], although we may never have heard it before. In the majority of cases we can not say with certainty which words we have learned from other people and which we have produced ourselves; in the majority of cases, as in the above cited examples, parallel forms make it possible to produce only one form, regardless of who is producing it. For this reason W. von Humboldt early on pointed to the perpetual creativity of language. (Kruszewski, 1995)

Kruszewski here suggests that the production of a novel inflected form for the Russian verb *vesti* 'lead' is guided by knowledge of other forms of *vesti* as well as other inflected forms of the different verb *idti* 'to go'. This represents, according to him, a clear example of an essential challenge presented to theory for language analysis, namely, the "perpetual creativity of language."

Familiar structuralist linguistic theories have operated with a misleadingly 'combinatoric' conceptualization of parts and wholes: wholes are of theoretical interest to the

[3] See Blevins (2016) for a detailed review of this tradition and its modern development under the label of Word and Pattern Morphology.

degree that they permit the identification of parts which can be recombined algebraically to recompose them, with little or no remainder. The whole as representing a distinct level of analysis is foreign to this conception, but is central to efforts to understand systemic organization: the internal structures of wholes serve to discriminate wholes from one another and the networks of relatedness patterns defined by these wholes constitute the analyzable organization of the system.

Significantly, this latter tradition, which developed in parallel with the more familiar Post-Bloomfieldian structuralist, morphemic approach,[4] displays conceptual and analytic affinities with research in the "developmental sciences"[5], where the fundamental constructs guiding explanation include "complex adaptive systems", "systemic organization", and, more generally, a focus on describing and understanding the dynamic interplay between parts and wholes on different interdependent levels that both constitute and define the organization of systems in both nature and culture.[6]

Segueing to the second question (4.1.30b) concerning the types of explanation invoked to account for E-complexity differences across languages, work in (typological) sociolinguistics has hypothesized that such differences may correlate with aspects of social structure: languages spoken by **large, diverse** populations are claimed to be morphologically **simpler** than those spoken by **small, close-knit** ones (Kusters, 2003; Perkins, 1992; Thurston, 1987, 1992; Trudgill, 2009, 2011, 2016; Wray & Grace, 2007). Adopting the terminology of Wray and Grace (2007), we refer to the former as EXOTERIC SITUATIONS and the latter as ESOTERIC SITUATIONS. See Table 4.1 for a summary of the characteristic properties of each.

We will contrast two basic categories of explanations about the relationship

[4]See Embick (2015) for a detailed discussion and defense of this "piece-based" conception of morphology.

[5]See Moore (2006) for an overview.

[6]See Ackerman and Nikolaeva (2014); Corning (2018); Hood, Halpern, Greenberg, and Lerner (2010); Jablonka and Lamb (2014); Laland (2018); Oyama, Gray, and Griffiths (2001).

Table 4.1: Comparison of esoteric and exoteric situations.

Property	**Esoteric situation**	**Exoteric situation**
Total community size	smaller	larger
Adult language contact	lower	higher
Learner population	primarily children	contains significant number of adults
Social stability	higher	lower
Communally-shared information and traditions	higher	lower
Morphological correlate	higher E-complexity	lower E-complexity

between the esoteric vs. exoteric state of a speech community and the E-complexity of its morphology. The first category of explanation can be referred to as *adaptationist*.

(4.1.31) Correlations between social and linguistic types are a matter of adaptation: some language types are 'fitter', and therefore selected for, in certain social environments.

Amundson (1996, p. 25), in developing a more catholic conception of explanation in evolution, identifies the adaptationist strategy as a primary informing hypothesis with a long history:

> To be sure, adaptationists admitted that organs and body parts exist which have no known adaptive purpose. The universal stance on these items might be called the principle of *presumptive adaptation*: Never infer a lack of adaptation from the lack of knowledge of adaptation, because it is always more probable that an unknown adaptive purpose exists than that no purpose exists. The presumption should be that the trait is adaptive, and that eventually its purpose would be discovered.

The primary exemplar of this category of explanation we will consider here is the LINGUISTIC NICHE HYPOTHESIS (Dale & Lupyan, 2011; Lupyan & Dale, 2010, 2015, 2016b):

> [L]anguages adapt to the learning constraints and biases of their learners. (Dale & Lupyan, 2011)

That is, the adaptationist explanation for the observed relationship between social structure and E-complexity is that **both** morphological simplification **and** complexification reflect adaptation to the different learning capacities of L2 and L1 learners in different social situations. As Lupyan and Dale (2010) put it:

> Our findings indicate that just as biological organisms are shaped by ecological niches, language structures appear to adapt to the environment (niche) in which they are being learned and used. As adults learn a language, features that are difficult for them to acquire, are less likely to be passed on to subsequent learners. Languages used for communication in large groups that include adult learners appear to have been subjected to such selection. Conversely, the morphological complexity common to languages used in small groups increases redundancy which may facilitate language learning by infants.

It is important to emphasize that an adaptive explanation is compatible with three hypotheses: it could explain both simplicity and complexity, only simplicity, or only complexity. It could also, of course, extend to none of these alternatives.

This perspective more broadly embraces a popular and previously prevailing analytic stance concerning the role of external forces on the modification of existing structures. Amundson (2005, p. 127) refers to this as *The adaptive rule of reconstruction* and formulates it as follows:

> The adaptive rule of reconstruction: Identify ancestral characters and selective forces such that the forces might have caused populations that possessed the characters to diverge into the descendent forms.

In effect, this strategy, characteristic of the Modern Synthesis in biological evolution, has been adopted in other fields which attempt to explain observable change in evolutionary systems: the operative notion is that factors external to the object of change both motivate and shape that change.

The second class of explanations we consider concerning the relation between complexity and social conditions is NEUTRAL:

(4.1.32) Independent of any forces of selection, random variation (evolutionary 'drift') can cause E-complexity to increase.

Existing examples of such explanations for sociolinguistic typological patterns can be found in Lass (1997a), Ehala (1996), Trudgill (2016), and Kauhanen (2017), *inter alia*. In the more general context of evolutionary systems, one formulation of this kind of explanation is offered by McShea and Brandon (2010):

> . . . in an evolutionary system in which there is variation and heredity, there is a tendency for diversity and complexity to increase, one that is always present but may be opposed or augmented by natural selection, other forces, or constraints acting on diversity or complexity.

That is, with respect to the observed correlation between social structure and E-complexity, increasing complexity may be the default state of evolutionary systems. This means that no additional explanation is necessary to account for increasing E-complexity in a given language, beyond whatever other contingencies may obtain. In this connection it is important to observe that while increasing E-complexity may have default status, the particular organization of the resulting system, i.e. its I-complexity, may be guided by both internal properties of particular systems as they co-evolve in conjunction with learnability considerations. In other words, the factors responsible for elaboration or simplification in E-complexity may be quite different from the factors responsible for the emergent organization associated with I-Complexity. This point is compellingly illustrated in Sims and Parker (2019) where it is shown that the mere enumeration of elements constitutive of e.g., Russian's inflectional morphological system does not provide insight into the important dimension concerning how these elements cohere, let alone, why they might cohere in the ways that they do. They conclude (p. 30):

> This suggests that the system as a whole is not simply a function of the complexity of its parts. It is instead a product of the way the parts are distributed –

> i.e., how the component elements are related. This should hardly be a surprise, but the data in this paper highlight that these sorts of local relations, and how they lead to complexity in an inflection class system (or don't!), are at least as important to focus on as the complexity of the system overall. To the extent that languages universally or predominantly exhibit low systemic complexity, the question becomes why. At a broad level, the answer likely has to do with learnability (Ackerman et al., 2009), but to get beyond general formulations of this idea, it will be necessary to dive into the learnability of specific inflection class configurations, and to carefully examine local relations among the component parts of individual inflection class systems.

This can be interpreted as suggesting the importance of distinguishing between E- and I-complexity: E-complexity as derivable from the World Atlas of Linguistic Structures (WALS)[7] provides inventories of morphosyntactic distinctions and their formal exponence, but these alone are simply the ingredients that get organized into the language particular systems that distinguish a language's morphological organization, i.e., I-complexity. Thus, any hypotheses concerning the relative influences of neutral or adaptationist factors need to clearly identify the scope of influence with respect to E- and I-complexity. For example, it may be that neutral factors influence the E-complexity of a language, while the organization of the resulting elements arises from some adaptationist considerations such as learnability, as mentioned in the preceding quotation.[8]

In this connection, it is important to observe that neutral explanation, and more generally, non-adaptationist perspectives, can be seen as complementing, rather than replacing adaptationist speculations about specific developments, and can themselves be seen as guided or biased by the internal dynamics of the specific systems (Arthur, 2004; Riedl, 1977; Whyte, 1965). Amundson (2005, p. 127) argues that this system internal perspective

[7] See §4.3.1 for discussion about the limits of what kinds of questions WALS can usefully address.

[8] It is also worthwhile in this connection to consider the valuable reflections contained in Chapter 10 of Bentz (2018). Of particular interest is the recognition that esoteric situations are often characterized by multilingualism, so that contact conditions and the influence of second language learning associated with exoteric situations is not necessarily associated with morphological simplification, as discussed in Meakins, Hua, Algy, and Bromham (2019).

on change and possibility for novelty, is the source of fertile reappraisals of adaptation as the single factor of change. He refers to the basic strategy as *The generative rule of reconstruction* and formulates it as follows:

> The generative rule of reconstruction: Identify an ancestral ontogeny that can be modified into the ontogenies of the descendent groups.

What is crucially distinctive here is the focus on ontogenies of development: the mechanics of how a system is organized and operates to yield effects over time.

In sum, the alternatives of adaptive and neutral explanation (with the latter supplemented by considerations of internally guided possible trajectories of change) provide the broader context of competing explanatory resources: while the former is often functionalist in nature, the latter is structuralist, following the traditional distinctions delineated in Amundson (2005).

Our aim in this chapter is to convey to the reader the nature of a neutral explanation of an evolutionary system's state and trajectory, and to convince the reader that this type of explanation is a strictly simpler and more likely explanation of higher E-complexity in esoteric situations than the Linguistic Niche Hypothesis. To accomplish this, we review in the next section the three defining properties of Darwinian evolutionary systems, why language change qualifies as one, what neutral vs. adaptationist explanations for the behavior of an evolutionary system are, and why the Linguistic Niche Hypothesis is adaptationist. In the third section, we begin by discussing the methodological challenges facing evolutionary explanations in biology, language change, and specifically the relationship between high E-complexity language variants and esoteric communities: a lack of data and a wealth of logically possible explanations with unclear or plausibly overlapping predictions. We argue that addressing these problems requires clearly (preferably mathematically) specified models of hypothesized causal mechanisms (e.g. learning), serious consideration of neutral hypotheses and evidence for them, and that simpler explanations (which will often be

neutral) be accepted over more complex ones by default. In the rest of the section we offer two such simpler (neutral or more neutral) explanations for the same phenomena as the Linguistic Niche Hypothesis. First, we point out more that the main independent variables of the Linguistic Niche Hypothesis — population size, structure, and other demographic parameters — have been known for more than a century to critically affect the relative likelihood of neutral vs. adaptive explanations of the state or trajectory of an evolutionary system; in particular, at least one neutral force — drift — is substantially stronger in small populations than in large ones, can easily be strong enough to overwhelm selectional factors identified in adaptationist approaches, and should be expected to lead to small populations exhibiting and maintaining traits that, if present in an otherwise identical larger population would be expected to disappear. Second, we review some recent literature modeling language change as an evolutionary process that investigate (among other things) the effects of social structure on the propagation of harder-to-learn vs. easier-to-learn linguistic variants. Together, they suggest that even when there is selection against a linguistic variant (i.e. uniformly for all learners in both more esoteric and more exoteric populations), the structure of esoteric vs. exoteric populations could lead to a relative homogeneity of input to learners in esoteric situations — enough homogeneity that linguistic variants that need more observations to be successfully learned are plausibly more likely to arise and persist in esoteric populations than exoteric ones. These two results mean that in the absence of strong (forthcoming) evidence for an adaptive explanation of higher E-complexity in esoteric situations (e.g. a benefit to L1 learning), neutral factors are both simpler and specifically more likely than adaptive ones to explain observations about the evolutionary trajectories of historically small, esoteric populations.

4.2 Background

In this section, we review the defining properties of Darwinian evolutionary systems and why natural language qualifies as one, offer a slightly more technical exposition of the difference between neutral vs. adaptive explanations (with examples from both biology and natural language), and then position the Linguistic Niche Hypothesis (henceforth LNH) with respect to these alternatives.

4.2.1 Darwinian evolutionary systems

A Darwinian evolutionary system can be defined in terms of three abstract elements (adapted from Lewontin 1970, 1978):

(4.2.33) a. A population of replicators: A population of objects capable of replicating themselves more faithfully than not from one timestep to the next.

b. Variation: Objects in the population can have potentially distinct traits along one or more dimensions.

c. Selection: Some variants in a population are better at replicating than others *by virtue of* differences in traits.

A trait value that causes those replicating objects that have it to display higher expected success at replication than those with some other variant of the same trait is 'adaptive'. Insofar as a trait is adaptive with respect to a particular kind of external environment that an object exists in or there is some internal aspect of the object's replication process that shapes its expected success at replication, that trait is said to make the object 'fitted' or 'adapted' to its environment or 'life cycle'.

Mechanistically, a Darwinian evolutionary system can be defined by a population state at some moment in time — a frequency or probability distribution over a set of variant types — and an algorithm by which the population at the next time step is generated from

the current one — i.e. a set of mechanisms or processes (in parallel or in some sequence) by which replication occurs. Replication involves two basic types of probabilistic choices: choosing for each object whether it survives and replicates, and for each of those that do, choosing how many copies result and how accurately those copies reflect the originals. A replication process that affects which objects survive or replicate only contributes to creating *variation* when the probability that an object is chosen for survival and replication as a result of that process doesn't depend on their variant type. Similarly, a causal mechanism that affects the number or accuracy of copies of an object chosen to replicate is a mechanism of variation if it doesn't depend on the variant type of the object. In contrast, a causal process affecting a population's dynamics is a *selection mechanism* when its effect on an object's probability of survival, probability of replication, expected number of copies, or the accuracy of those copies depends on the variant type ('traits') of the object.

In the context of biology, examples of different kinds of populations of replicators include

(4.2.34) a. populations of alleles — different values or forms of a gene

b. populations of genotypes — different partial or complete genomes

c. populations of phenotypes — different combinations of physical and behavioral traits of an organism.

The question of which of these is the most appropriate 'unit of selection' can depend on theoretical commitments about biology or evolutionary theory, what scientific question is being addressed, what method has been chosen, or what data are available. Examples of variation mechanisms include

(4.2.35) a. random choice of which organisms die and which reproduce independent of each organism's variant type ('drift', discussed in the next subsection)

b. random mutation of alleles during replication

c. random migration to or from other populations.

Some examples of ways that variant types can differ in terms of fitness (i.e. selection mechanisms) include

(4.2.36) a. probability of survival (viability selection)

b. probability of reproducing (e.g. sexual selection — the probability of finding a mate)

c. expected number of offspring per reproductive event (fertility selection).

An important type of selection that cross-cuts classification by biological life stage — and is particularly relevant to cultural evolution — is *frequency-dependent selection*, where the fitness of an individual with a given trait is a function of the relative frequency distribution over traits in the population; we discuss this more below in the context of language. Finally, note that in biology every variation and every selection mechanism listed here is capable of causing the frequency distribution over traits to change, and every variation and every selection mechanism can cause a trait to *disappear* from a population, but only some variation mechanisms (e.g. mutation or migration) can *introduce* a previously absent trait and only some variation and selection mechanisms can act to *maintain* variation within a population.

While the three abstract elements of (4.2.33) suffice to define a Darwinian evolutionary system, in both biology and language, populations of replicators typically have *structure* that affects what the replication process is, how variation is introduced, and how selection filters or amplifies variation in ways that are substantive, scientifically interesting, and particularly relevant to discussion of the relationship between demographic factors and the relative effects of drift vs. selection. That is, a population is supposed to represent a set of spatiotemporally-bounded and co-occurring individuals that live, compete, cooperate, and reproduce together in the same context. Suppose a population of individuals (modeled

or empirically observed) is meaningfully dividable into two or more subpopulations with a limited and potentially non-uniform rate of migration between them — e.g. subpopulations of an organism may be subdivided over different social groups (herds, flocks, etc.) and/or multiple locations like isolated meadows or lakes, an island and a mainland, or the islands of an archipelago. If we want to model the dynamics of this population, we can incorporate our beliefs about this subdivision and organization of the population as accurately as we can, or we can idealize over these differences and treat the population as though it were less structured; our motivation may be practical — a lack of data or the desire for a more analyzable model — or theoretical — e.g. exploring how much or little population structure affects the dynamics of the whole population and each of its subpopulations. We roughly summarize the effects of population structure below:

(4.2.37) a. All else being equal, the lower the rate of migration, the less the dynamics of each subpopulation are affected by others.

b. The more asymmetric and heterogeneous population sizes, forces of selection, and migration rates are between populations, the more inaccurate it will be to lump the subpopulations together and treat them as a single unstructured population in a single environment.

c. The higher the average rate of migration, the more symmetric migration is between subpopulations, and the more similar population sizes and forces of selection are across subpopulations, then the more accurate of an approximation it will be to treat this 'metapopulation' as a single unstructured population.

We elaborate in the next section on the relationship between population structure, population size, the effects of forces of variation vs. selection, and implications for the LNH.

Analyzing language as an evolutionary system involves making several choices. We schematize these choices as follows:

(4.2.38) a. *Choosing the set of linguistic representations that variants will be drawn from.* For example, variants could be different pronunciations of a phoneme, different strategies for expressing a morphosyntactic property, different synonyms for a meaning, different grammatical strategies for encoding a meaning, generally all or part of a grammar concerned with defining 'different ways of saying the same thing' (Croft, 2000, p. 31), or distributions over any of these choices of a set of variants.

b. *Choosing a replication timescale — individual dyadic communication episodes vs. language development.* At its most granular, 'replication' can be taken to be the production of a unit of form — possibly with some meaning and in some episode-specific context — followed by the recognition or comprehension of that form by a listener and some update of the speaker and listener's representations of what the language is. Alternatively, replication can correspond to an abstract (child) language development event where some speaker-teachers of the existing community are chosen to provide the input to a learner, who then chooses a linguistic variant (or distribution over variants) at the end of the process and becomes a new speaker-teacher member of the population at the next timestep.

c. *Choosing a relationship between the population of linguistic variants and the population of speakers in a speech community.* The basic replicating object can be taken to be a token of a linguistic variant, and each speaker in a speech community at time t can be associated with a population of such tokens — interpretable as a distribution of remembered observations (e.g. 'exemplars') and/or a production distribution over variant types, and a speech community then corresponds to a population of subpopulations (a 'meta-population'). Alternatively, the basic replicating object can be identified with a speaker and

their linguistic representation — e.g. a single linguistic variant, a grammar, or a distribution over variants — and a speech community at a particular point in time can be treated as a population.

While the first choice is relatively straightforward, the last two are more complex and interrelated. For the purpose of understanding language change as an evolutionary process, we discuss different combinations of options for these last two choices below and sketch what population structure and variation vs. selection mechanisms look like under each such choice. We begin with the most granular choice of timescale and population.

The most fine-grained choices of replication timescale and population take each speaker in a speech community to represent a population of linguistic variant tokens, the speech community to represent a meta-population, and individual dyadic communication episodes to be the main process by which the distribution of variants changes over time. Each speaker's population of tokens is most plausibly interpretable as a set of variant tokens or distribution over variant types representing what that speaker has observed themselves and others produce to date,[9] or some function of such a distribution (Blythe & Croft, 2012; Reali, Chater, & Christiansen, 2014; Wedel & Fatkullin, 2017; Winter & Wedel, 2016). Replication principally involves repeatedly choosing a speaker and listener pair who will interact, choosing what the speaker says, and an update process describing how one or both participants adjust their internal distributions over linguistic variants as a result. Below is a sequence of events describing how this interaction and update process could be modeled for some choice of speaker s and listener l.

(4.2.39) a. Suppose there are $X = \{x_1 ... x_k\}$ different types of linguistic variants, and that the speaker has to date observed $O_s = \{o_1, o_2 \dots o_i \dots o_n\}$ tokens, with the

[9]Note that these could be taken to be perfect or lossy representations of such observations; if they are lossy representations, then the lossy compression and/or noise process by which observations are modified is part of the replication process.

variant type of o_i given by $v(o_i)$.

b. Based on the speaker's observations O_s and a learning or inference algorithm L, the speaker currently has a production distribution $p_s(X|L(O_s))$. They choose a single form x^* to produce by sampling from p_s. A simple example production distribution — exhibiting no selection — might have them randomly choose one of their past observations: $p_s(X = x^*) = n^{-1} \cdot |\{o_i \in O_s | v(o_i) = x^*\}|$.

c. The speaker produces a token x^* and adds it to their set of observations.

d. The listener perceives the actually produced form as y, where $p_n(Y = y|X = x^*)$ describes how noise can cause the listener to perceive y as something different from x^*.

e. The listener arrives at some beliefs $p_l(X = \hat{x}|Y = y)$ about what the speaker actually produced. For example, the listener might reason Bayesianly by combining y with a prior model of what the speaker is likely to have intended to produce $p_s^{'}$ and a model of the noise distribution p_n as $p_l(X = \hat{x}|Y = y) \propto p_n(Y = y|X = \hat{x})p_s^{'}(X = \hat{x})$.

f. Using this distribution $p_l(X|Y = y)$, the listener chooses some estimate $\hat{x}$ according to a decision rule — e.g. choosing the $\hat{x}$ that maximizes $p_l(X = \hat{x}|y)$ — and adds it to their own set of observations.

In sum, production by the speaker causes a token of some linguistic variant x^* to replicate in a 'population' of observations associated with the speaker, and after potential modification by noise, perceptual/comprehension processes, and a learning process, to replicate in a 'population' of observations associated with the listener.

To permit examination of meaning and form-meaning mappings, this variation of the scenario could be slightly modified by starting each interaction episode with a randomly chosen meaning (interpretable as e.g. a uniquely salient referent in the common ground) that

is in the common ground and that the speaker uses in choosing a variant form to produce; by combining this observation of the co-occurrence of meaning and form with existing beliefs about this dimension of the language, the listener can adjust their distribution over form-meaning mappings.While this scenario permits investigation of form-meaning mappings, we can extend it further to investigate comprehension and the effects of communicative success or failure on how interlocutors adjust their distribution over variants at the end of each episode. Instead of having the listener jointly observe a uniquely salient meaning and the speaker's produced form, the speaker can privately choose a meaning from a common-ground set of meanings (e.g. a physically salient set of referents) and attempt to communicate it to the listener. After perceiving the speaker's produced form variant the listener can then reason about what meaning the speaker intended and do something — e.g. 'point' at a particular referent or take an action based on their beliefs about the speaker's intended meaning — that conveys information to the speaker about the listener's interpretation and whether the speaker was successful. One or both participants can then update their beliefs based on each other's existing beliefs and observed behavior. Finally, to explicitly represent heterogeneity in types of learners — e.g. children vs. native adults, contact between language varieties, L2 learners, etc., each with e.g. some different initial distribution over observations or learning process — we can specify a rate at which a speaker-listener is added or removed to the population, and a distribution over *what kind* of speaker-listener is added or removed.

Forces of variation and selection here are determined by

(4.2.40)
a. the probability distribution over which pairs of individuals are chosen to be speaker-listener pairs
b. the probability distribution over what a speaker intends and actually produces
c. how production of a token affects the speaker's population of variant tokens
d. the probability distribution over what a listener perceives and/or comprehends

given what the speaker produced

e. how a listener's beliefs about what the speaker said and/or meant affects the listener's adjustment of their population of variant tokens.

f. any other details about memory and inference process specifying how observed tokens of linguistic variants are stored and shape future inference and decision-making of a speaker-listener.

That is, if the probability that any pair of individuals are chosen to be speaker and listener does not depend on the variants of the pair (or distributions over variants of the pair), then that aspect of the replication process would contribute to variation but would not involve selection; similarly, if what the speaker produces, how accurately it is produced or perceived, or how it affects a listener's future inference or production behavior does not depend on the variant of the speaker or listener, then those aspects of replication contribute to variation but would not involve differential selection of some linguistic variants over others. An example where this process would involve selection is a model where some linguistic variants are associated with different social groups or identities and speakers or listeners are more likely to be paired with an interlocutor whose variant (or distribution over variants) is relatively similar[10] — or relatively different — with respect to social dimension. Another example illustrating frequency-dependent selection: suppose a speaker is capable of producing two or more variants that differ greatly in how prevalent they are among the speech community the speaker interacts with, and where many other speakers are unlikely to be familiar with or understand the rarer variants (the variants could e.g. be different wordforms for the same meaning associated with different language varieties). If the speaker is aware of this difference in relative frequency and it causes them to prefer to produce the more frequent variant more often than they otherwise would as a result, then the rate of

[10]This would be a sociolinguistic analogue of *assortative mating* where organisms preferentially mate with others that are similar to themselves.

reproduction (fitness) of variants is a function of the frequency of different variants and will lead — in the absence of forces or conditions like drift, opposing forces of selection, or population structure creating conditions to maintain rarer variants — to a 'rich get richer' dynamic where whichever variant is more frequent will become even more frequent.[11] Other examples of ways in which variants could be differentially selected include the following:

(4.2.41) a. Some variants may be more likely to be misheard (J. Ohala, 1993) or misunderstood by listeners, or be more likely to vary or be misproduced by speakers.

b. If speakers and listeners have distributions over linguistic variants, then a speaker may preferentially produce some variants over others if they vary in terms of their estimated sociolinguistic utility (signalling e.g. group identity or prestige), or in terms of their estimated communicative utility (in the sense of e.g. Lindblom, 1990). This production preference over variants could depend on the speaker's own distribution, the speaker's model of the listener, other communicative and social aspects of the situation, or generalizations the speaker may have made from past experiences, including e.g. their estimate of the probability distribution over variants of other individuals in the speech community. Note that most of these possibilities are examples of frequency-dependent selection.

c. Listeners may differentially weight or discount a speaker's produced variant in updating their own linguistic variant or distribution on variants in a way that depends on the speaker's produced variant or the listener's variant. This could be caused by e.g. the sociolinguistic properties of the variant, the listener's distribution over variants, or from the listener's estimate of the distribution among other individuals in the speech community. Again, some of these possibilities are examples of frequency-dependent selection.

[11]Note that frequency-dependent selection could also go in the other direction: a preference for *novelty* will lead — all else being equal — to differential replication that favors variants that are *rare*.

Note that as long as the conditions in (4.2.33) are satisfied, we have a Darwinian evolutionary system: no one choice of replicator or timescale of replication here is necessarily exclusive with another. In fact, the model setup and mechanisms described above can be interpreted at a coarser level of analysis, where the timescale of replication is still dyadic communication episodes, but the population of interest (in the sense of (4.2.33)) is taken to consist of entities (speaker-listener distributions over variants) that *happen* to also be interpretable as populations. (Hence the term 'meta-population'.) Here the space of variant types consists of the space of possible speaker-listener states (the space of distributions over linguistic variants) and the replication process describes how each speaker-listener's population state changes after a communication episode, exactly as before.[12] As before, if the probability that two members of the speech community are chosen to interact as speaker and listener depends on their variant types, then that would constitute a selection mechanism. Similarly, if some variant types (population states) are more likely to accurately replicate than others, then that would also be an example of a selection mechanism.

Coarsening the replication timescale, the replication process can instead abstractly describe (child) language development. This involves choosing a set of one or more speaker-teachers from the set of current speakers, and based on that choice — e.g. by sampling data from each teacher and applying a model of learning — generating a new speaker with a new linguistic variant or, more generally, distribution over variants. As before, if each speaker is associated with a probability or frequency distribution over linguistic variants, each speaker can be interpreted as a population of linguistic variants and each speech community as a meta-population, or (equivalently) a speech community can be interpreted as a population whose members are distributions over linguistic variants. The essential differences from the previous choice of timescale is that adults are modeled as static, children learn only from

[12]Learning and inference correspond to replication in the sense that e.g. a speaker-listener's updated variant distribution at time $t + 1$ after an interaction at time t is a function of the distribution at time t.

interactions with the previous generation of adults, and details necessary for specifying the process and outcomes of dyadic communication episodes can be abstracted over.

An example of work following this schematization of language change as an evolutionary system includes work on *iterated language learning* (S. Kirby, 2001).[13] This is a relatively simple model of cultural evolution intended to facilitate investigation of how the cumulative effect of mechanisms of cultural transmission (i.e. learning) can shape cultural conventions like language over the course of many generations. In the basic version of this model (Griffiths & Kalish, 2007), each 'generation' consists of one learner. Each agent in a generation learns by observing a sample $O = \{o_1 \dots o_n\}$ of the cultural behavior (e.g. a set of forms or form-meaning pairs) of the previous generation and then Bayesianly updating their prior beliefs $p(G)$ about what the most likely causes (e.g. underlying grammar(s) or lexicons) of the data they observed are: $p(G|O) \propto p(O|G)p(G)$. Each agent then samples a hypothesis (grammar and/or lexicon) g from their distribution over causes $p(G|O)$ and proceeds to produce data according to their chosen grammar or lexicon for the next learner generation, i.e. according to $p(O|g)$.[14] The prior over grammars $p(G)$ reflects the inductive biases of learners; all else being equal, it determines which hypotheses are easier or harder to learn.[15]

The simplicity of this basic form and the use of a Bayesian model of individual inference permits laboratory experiments (see Irvine, Roberts, & Kirby, 2013; S. Kirby, Griffiths, & Smith, 2014; Mesoudi & Whiten, 2008, for reviews and critical evaluation), extensive mathematical analysis (e.g. Griffiths & Kalish, 2007) of model behavior and experimental results, and separate manipulation of linguistic representations, population

[13] See also earlier work by Esper (1925, 1966).

[14] There is typically also a small, fixed, and uniform probability of making a production error.

[15] The more data is available, (averaging over possible sets of observations) the less a learner's prior matters and the closer their posterior $p(G|O)$ will be to the distribution with all mass concentrated on the teacher's actual chosen grammar. See Griffiths and Kalish (2007, §3.1).

structure, and processes of production, comprehension, and learning. Finally, note that Reali and Griffiths (2010) establish a general correspondence between parameter values for a variant of the basic iterated language learning model and the mutation rate of the Wright-Fisher model with drift and K alleles (generalizing beyond the value of $K = 2$ illustrated in the previous section). This result offers a mathematically explicit bridge for connecting the large body of literature on biological evolution to work on iterated learning and forcefully suggests that the arguments about the explanatory burden of neutral vs. adaptationist models offered in §4.3 rest on more than just an analogy or abstract similarity between biology and language.

4.2.2 Adaptive vs. neutral explanations of variation

Given an evolutionary system, what scientific questions can we ask about it? As summarized by Stephens (2008),

> Population genetics is the study of processes that influence gene and genotype frequencies. It has been obsessed with two related questions: what is the extent of the genetic variation between individuals in nature and what are the factors that are responsible for this variation?

The two questions Stephens identifies apply to any evolutionary system, and answers to them generally emphasize one of (4.2.33b) or (4.2.33c) more strongly than the other: *neutral* explanations emphasize the role of mechanisms of variation, while what we have termed *adaptationist* explanations focus on mechanisms of selection. The question of which type of force is more important (and in what sense) to explaining the extent and dynamics of an evolving population is one of the oldest and most important debates in evolutionary theory.

Below, we exemplify neutral processes in both biology and language: we introduce one of the basic models of population genetics (discussed in more detail in the next section) where a neutral process (drift) is by hypothesis the only force affecting the dynamics of

the population, and we discuss an empirical example of complexification in morphosyntax without any obvious or likely explanation in terms of selection.

One of the strongest examples of an answer emphasizing variation mechanisms in biology is 'neutral theory' (Kimura, 1983), which holds that at the molecular level,[16] mutations and variation we observe are fitness-neutral (or nearly so) and that any given variant's apparent ubiquity within a population (the 'fixation' of a particular variant and the disappearance of alternatives) is more likely a consequence of *drift* than selection. Drift models the fact that sometimes an organism (or instance of a gene, etc.) in a generation is replicated more or less often than others in the same generation as a result of *chance* rather than another neutral process — like migration from another population — or a form of selection. That is, drift is one of the simplest ways in which a population of imperfect replicators can imperfectly replicate: a completely random subset of the population is chosen for replication (some potentially more than once), and the rest fail to replicate at all. Fig. 4.1 illustrates the hypothetical trajectory of a very small constant-size population ($n = 10$) of gametes of asexually reproducing organisms, where each organism is an instance of one of two possible variants — 'blank' or 'filled'. The generation at time t_{i+1} is created by sampling with replacement n times from the generation at time t_i: these samples are the members of the new population. Drift is a neutral process because the probability that any particular member of the population at time t_i will replicate doesn't depend on or differ based on the traits of that individual. If, instead, one variant was explicitly more likely to be chosen for survival and replication than the other, then the population would be evolving under both drift and selection. To foreshadow discussion in the next section, observe that even though the population in Fig. 4.1 started evenly split over both variants, it is quite likely that the population will end up consisting entirely of the 'blank' variant within just a few timesteps of t_4 — a complete change in the trait diversity of this population in a handful of generations,

[16] I.e. as opposed to the genetic — a 'gene' is an abstraction over molecules.

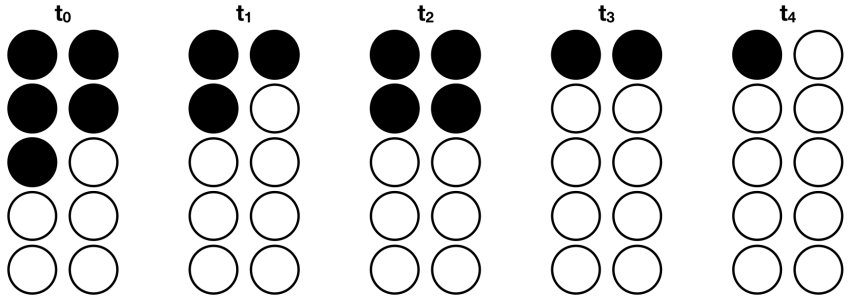


Figure 4.1: A graphical illustration of drift acting on a small population with two variants.

all without any forces of selection. In the next section we elaborate on the interplay of drift, population size, and selection, and discuss implications for adjudicating between neutral vs. adaptationist explanations of variation and the Linguistic Niche Hypothesis.

A linguistic analogue of neutral changes and processes is offered by Trudgill (2011, 2016), who discusses an example from a traditional dialect of southwestern England that underwent morphosyntactic complexification without any obviously adaptive explanation. In this dialect, intransitive infinitives became marked with a word final morpheme **-y**, yielding the type of general transitive versus intransitive contrast schematized in (4.2.42a). The actual encodings are presented in (4.2.42b) and (4.2.42c), where the infinitival form required after the modal 'can' is either affixless for transitives, or affixed by a **-y** for intransitives.

(4.2.42) a. *to hit* vs. *to runn**y***

b. Can you zew up thease zeams?
Can you sew up these seams?

c. There idden many can shear**y** now.
There aren't many who can shear now.

According to Trudgill (2011, 2016), this typologically unusual marker likely arose as a reanalysis of a phonologically conditioned change. That is, before this innovation arose, all Middle English infinitives had an ending — [-i] for the relevant southwestern dialect variant. We also know that eventually this word-final unstressed vowel was lost in almost all dialects. Appealing to observations of analogous ongoing variation and change in Scandinavian dialects, Trudgill suggests that before this loss was complete, there was a period of variation during which some types of infinitives were more likely to lose this vowel slower — or faster — than others, namely utterance-finally vs. between words (i.e. before obligatory object nouns). While this infinitive marker eventually disappeared everywhere in other dialects, speakers in this southwestern dialect reanalyzed phonologically-conditioned variation as an obligatory morphosyntactic marker.

There is no salient reason to think that speakers preferentially produced or learners preferentially inferred — during this transitional period and in this location in England, but very few other similar contexts — a grammar with explicitly marked intransitive infinitives. Consequently, the propagation and survival of this convention in this speech community is most parsimoniously explainable in terms of neutral processes alone — one or more initial speakers inferred a morphosyntactic reanalysis of phonologically-conditioned variation, began producing data consistent with that reanalysis, and other speakers followed suit; eventually it became a convention of that speech community.

4.2.3 The Linguistic Niche Hypothesis

With a clear sense of the scientific question at hand and two categories of answers, we can now spell out in more detail what makes the claims of Lupyan and Dale (2010, 2015, 2016b) and Dale and Lupyan (2011) about the relationship between social situation and morphological complexity adaptationist. The LNH's predictions and the chain of reasoning

behind them (Lupyan & Dale, 2015) are summarized below:[17]

(4.2.43) Predictions:

a. EXOTERIC condition: The **higher** the population size and the **more** area a speech community is spread out over, the *less* inflectional morphology its language is likely to have.

b. ESOTERIC condition: The **lower** the population size and the **smaller** the area a speech community is spread out over, the *more* inflectional morphology its language is likely to have.

(4.2.44) EXOTERIC linking hypotheses:

a. **Increasing** population size and the area the population is spread out over is associated with a relatively **higher** proportion of adult L2 learners.

b. A **higher** proportion of adult L2 learners means that there is a **smaller** portion of the population than there otherwise would be that is likely to be able to successfully learn and use more complex inflectional morphology, likely leading to a trend of **decrease** in the amount of inflectional morphology, *ceteris paribus*.

(4.2.45) ESOTERIC linking hypotheses:

a. **Decreasing** population size and the area the population is spread out over is associated with a relatively **lower** proportion of adult L2 learners.

b. A **lower** proportion of adult L2 learners means that there is no force selecting *against* the propagation of more complex inflectional morphology.

c. Insofar as inflectional morphology is redundant and a more accessible kind of cue for child language learning than e.g. social or pragmatic reasoning reliant on extralinguist context, linguistic variants with more inflectional morphology may

[17]It is important to keep in mind that the predictions below reflect the E-complexity properties discussed in §4.1: they ignore how these properties are organized in terms of I-complexity.

be learned faster and/or more accurately by children than a language variant with less, leading to a trend of **increase** in the amount of inflectional morphology in the language, *ceteris paribus*.

As schematized here, linking hypotheses (4.2.44b) and (4.2.45c) can be understood as describing esoteric and exoteric social situations as different *epistemic environments* where different types of linguistic variants propagate ('replicate', 'transmit', 'are learned') more accurately and/or easily by virtue of being more appropriately matched ('adapted') to the strengths and weaknesses of the learner population: exoteric environments select against E-complexity and esoteric environments select for E-complexity. These are what makes the LNH an adaptationist explanation of morphosyntactic variation.

4.2.4 Interim Summary

Our goal in this section has not been to state the last word or offer definitive technical characterizations of either evolutionary systems generally or language specifically, but rather to illustrate for a linguistic audience the basic structure of an evolutionary process, a basic scientific question one can ask about such processes (viz. the relative burden of neutral vs. adaptive explanations), why language change meets the criteria of an evolutionary process, and why the Linguistic Niche Hypothesis is an adaptive explanation. That is, there are many subtle questions about evolutionary systems that are important to both theory and empirical measurement — e.g. What is the most appropriate unit of selection? When is a trait an 'adaptation'? What is the 'function' of a trait? — but not to our larger rhetorical goals in the next section: communicating basic results about the strength of selection vs. drift in Darwinian evolutionary systems as a function of population parameters like size, the difficulty of clearly identifying selection as the explanation for the distribution of a trait in a population, and why together these make neutral forces a more likely source of explanation for the linguistic typology of historically small speech communities, contra the

Linguistic Niche Hypothesis. We submit that language scientists interested in strong claims about language change as an evolutionary process should be aware of questions and debates in evolutionary theory and consult a textbook on population genetics (e.g. Hartl & Clark, 1997; Rice, 2004) or a survey of philosophy of biology (e.g. Hull & Ruse, 2008; Rosenberg & McShea, 2008; Sarkar & Plutynski, 2008).

4.3 The burden of evidence is on adaptive explanations

In this section we discuss two problems facing explanations of variation and change in evolutionary systems: data are generally few and expensive to acquire, and what data we have are often only weakly informative about which of many mechanisms (singly or in combination) caused them. We begin in the first subsection by considering the status of each of these two problems in biology and how it has affected the development and evaluation of theories and explanations there. We then proceed by considering whether similar challenges face the study of language change in general and the relationship between social situation and E-complexity in particular. We conclude that they do, and argue for three conclusions about theory development and evaluation for evolutionary explanations of language change:

(4.3.46) a. Evolutionary theories of language change need clearly specified models of hypothesized mechanisms affecting replication — e.g. learning.

b. Neutral hypotheses and evidence for them need to be considered and weighed alongside adaptive ones.

c. Simpler explanations should be preferred over more complex ones — especially in the absence of unambiguous data or explicit hypotheses with clear predictions. As discussed below, neutral models are often the simplest explanation.

In the particular case of the Linguistic Niche Hypothesis and the relationship between high

E-complexity and esoteric social situations, we point out that

(4.3.47) a. To date, no explicit model of the hypothesized mechanisms has been offered.

b. No or almost no effort has been expended on considering alternative hypotheses that would also predict a correlation between high E-complexity and esoteric social situations.

In the second and third subsections, we argue that there are simpler alternative explanations of a correlation between high E-complexity and esoteric situations that do not require there to be any selective pressures for high E-complexity in general or specifically in esoteric situations, and that therefore the burden of evidence on the LNH is even higher than previously appreciated. Specifically, in §4.3.2 we elaborate on how one of the simplest neutral evolutionary forces — drift — is significantly stronger in small populations than large ones, meaning that we should expect more typological variation across small populations than large ones and that whatever forces of selection are present in them will be blunted or plausibly even overwhelmed by the effects of drift. In the final subsection, we review two recent models of language change that consider (among other things) the effects of high vs. low diversity in the language variants of the initial speaker population and of esoteric vs. exoteric social network structures. Together, they suggest the relative homogeneity of input in esoteric social situations relative to exoteric ones means that any linguistic variant (e.g. potentially high E-complexity ones) requiring more observations to learn is more likely to be learned in an esoteric social situation than an exoteric one — crucially without any requirement that learners specific to the esoteric environment favor the more difficult variant or that learners specific to the exoteric environment favor the simpler variant.

4.3.1 Challenges of explanation in evolutionary systems

Stephens (2008) describes some of the challenges facing attempts to explain

variation in an evolutionary system and offers one of the key methods by which the study of biological evolution has made progress:

> ...Much of the historical, methodological, and philosophical interest in population genetics results from the fact that [its] two central questions —- the extent and explanation of genetic variation —- have proved extraordinarily difficult to answer. It is impossible to know the complete genetic structure of any species, and there are significant underdetermination problems in figuring out which factors are the relevant causes of evolutionary change, even if one knows a lot about the genetic structure of a population. Despite these difficulties, population genetics has had remarkable successes, and is widely viewed as the theoretical core of evolutionary biology. Significant evolutionary changes often occur over thousands or millions of years. Because of this, it is impossible to observe these changes directly. As a result, understanding the causes of evolution depends crucially on theoretical insights that flow from the mathematical models of population genetics.

That is, in the face of data about genetic variation that were both hard to come by and a variety of hypothesized mechanisms by which that variation could change (rendering most data underinformative), biologists expended great effort in elucidating the space of theories by constructing explicit mathematical models where the presence of different causal mechanisms affecting replication can be toggled on or off, parameters (e.g. population size, mutation rate, strength of selection) can be varied or related to empirical measurements, and the predictions of different modeling assumptions can be compared to each other and what data is available. As elaborated in the next subsection, these formalizations of Darwinian evolutionary dynamics show that drift should be expected to have a strong effect on the evolution of small populations and relatively little effect on large ones.

Mathematically explicit theories of evolution were not enough, however — they needed to be complemented by careful scientific reasoning about available evidence and consideration of available explanations. Historically, one of the main arguments of critics of adaptationist explanations in biology (prominently Gould & Lewontin, 1979) was that

researchers offering such explanations for empirical phenomena often failed to seriously investigate or consider the relative evidence for neutral explanations of the same phenomenon and accepted the apparent sufficiency of an adaptationist explanation on the basis of weak empirical evidence. Nevertheless, it is commonly noted (see e.g. Pigliucci & Kaplan, 2000) that one of the legacies of Gould and Lewontin (1979) over the last few decades has been an improvement in standards of evidence for adaptive explanations in evolutionary biology.[18]

In sum, in the face of insufficient empirical data and a complex hypothesis space full of theories making overlapping predictions, evolutionary biology proceeded in two directions: (1) clarifying mathematically the nature of each hypothesized causal mechanism affecting replication, identifying what data it predicts, as well as how it compares or combines with other mechanisms, and (2) by holding adaptive explanations of empirically observed variation and change to a higher standard of evidence.

What is the situation facing language? Generally speaking, data about variation and change is at least as hard to come by and at least as indeterminate with respect to ultimate causes. In fact, even our theories of causal mechanisms affecting replication and their relative frequency, strength, and interaction are in their infancy: insofar as we have explicit models of language learning, comprehension, or production in individuals, relatively little work has examined how these function at population- and historical-scales, how they interact, how or when each should be expected to be strong or weak, or how they relate to sociolinguistic factors. Finally, while we discuss some recent work in the next section that has begun to address these problems, few to our knowledge have yet examined detailed or realistically complex linguistic representations. Turning to evidential standards for adaptive vs. neutral explanations in language, Lass (1997a) argues that much functionalist work

[18]Note also that one of the important roles of mathematical models of the neutral theory of molecular evolution (Kimura, 1983) was providing a null model for inferring the presence and strength of selection from molecular data.

(including in the context of morphology) assumes that there is some teleological force of change in the direction of transparent (one-to-one) form-meaning mappings, motivated by the putative need to resolve the absence of clear function-form organization whenever this occurs. Discussing a representative proponent of this principle (Anttila, 1989) who dubs it *the mind shuns purposeless variety (*MSPV), Lass provides several examples where the principle appears to obtain, while demonstrating that there are many others where it does not. What is the status of such an adaptationist principle, given such a state of affairs? (Lass, 1997b, p. 344)

> If we invoke MSPV only for good outcomes, and allow bad ones to be not counterexamples but simply non-instantiations of something 'tendential' in the first place…the MSVP explanation is invincible, and therefore uninformative…This suggests that either 'the mind' doesn't behave this way (the obvious conclusion); or that the variety is not purposeless, or is at least neutral, in the sense that preference and dispreference are both arbitrary.

The lesson of present relevance for the analysis of language change is one of caution: adaptive explanations require careful elucidation in each instance and should be distrusted without such specification.[19]

Accordingly, while empirical data about language change and its mechanisms continues to accumulate, we argue that adaptive explanations of language change need to clearly describe hypothesized mechanisms and weigh evidence for their hypothesis with evidence for neutral ones. In cases where, on the one hand, data are sparse and relatively

[19]There is another lesson which applies to the general information-theoretic implicative framework which guides this chapter. Morphological organization analyzed in terms of low conditional entropy between words does not mean that such systems strive toward lower and lower conditional entropy values: in fact, conditional entropies can increase over time. Languages simply utilize whatever forms arise and (re)organize them into systems of greater or lower conditional entropy, as long as they retain enough transparency to be learnable. In other words, following Lass' observation, changes are not driven by tendentious (dis)preferences: there are, to our knowledge, innumerable (re)organizations compatible with the need to be learnable. Maiden's documentation (Maiden, 2018) of *morphological perseverance*, i.e. the maintenance of complexity where simplification would be expected owing to functional considerations, provides a challenge to change necessarily being motivated by impressionistic learnability considerations.

indeterminate, and on the on the other hand, we do not have a clear sense of what the space of hypotheses is or what predictions they make because mechanisms of language change have not or are only beginning to be explicitly formulated and analyzed, the principle of Occam's razor suggests that we should prefer simpler explanations over more complex ones. Insofar as neutral explanations of available data typically require fewer and/or weaker assumptions about what drives evolutionary change than adaptive ones do, they ought to be regarded as a priori more likely.

In the specific case of the Linguistic Niche Hypothesis's adaptationist hypothesis about E-complexity and esoteric communities, the situation outlined above for language with respect to data, theory, and consideration of alternative neutral explanations is even more pronounced. On top of the uncertainty about the relevance of E-complexity vs. I-complexity for understanding morphological complexity expressed in §4.1, it is still unclear that there is a veridical correlation between E-complexity and esoteric communities.

First, the statistical correlations proposed by Linguistic Niche Hypothesis and others are based on language data drawn from the World Atlas of Linguistic Structures, or WALS (Dryer & Haspelmath, 2013). WALS was constructed to support typological investigations by linguists. While it has proven its worth in that domain, many of its properties make it less well-suited for use in large-scale regression models. The information in WALS was collected over many years by various research groups for different purposes, and this naturally has led to large variation in quality and detail. Unavoidably, coding errors have crept in. For example, Rubino (2013) lists Nandi as a language which exhibits productive reduplication, but the source cited for this information is actually describing Kinande, an unrelated language. In many other cases, the coding is technically correct but obscures important differences between languages. Take the entry for number of nominal cases (Iggesen, 2013), a linguistic property that is clearly an aspect of E-complexity. The number of cases in a language would seem to be straightforwardly quantifiable as an integer.

But, in WALS it has been arbitrarily discretized into eight categories. This makes spatial visualizations simpler, but complicates the use of this feature in further statistical analyses. More deeply, individuating and enumerating cases is not without problems, even setting aside the issue of quantization. English is listed as having two cases; while this is not wrong, exactly, case marking in English is marginal at best and has a very different status in the grammar than it does in, say, Modern Irish. Also, cases with non-syntactic functions (like the vocative) were left out of the counts, as were genitives that agree in person or number with the possessed noun. These choices are justified and documented in the relevant WALS chapter, but subtleties like this get lost when many different features are combined into a single large statistical model.

Second, most of the demographic information we have is only weakly informative about the Linguistic Niche Hypothesis's hypothesis and its object of explanation: most measurements that we have are limited to recent history and languages with historically small speech communities are in general the ones for which we are likely to have the least data, especially historical data. Finally, the problems with each of these sources of data compound each other when correlations between them are examined: only some fraction of demographic data about a speech community is likely to be associatable with relevant historical descriptions of the language with enough detail to draw conclusions about E-complexity.

Turning to theorized mechanisms and empirical predictions, the LNH's hypothesis about the relationship between E-complexity and esoteric communities to date has little empirical data and no explicit models indicating

(4.3.48) a. why high E-complexity could or should be expected to facilitate L1 but not adult L2 learning,

b. why ease of learning among children of higher E-complexity variants is at the

expense or exclusion of later acquisition or use of lower E-complexity variants,

c. that this pressure *for* high E-complexity everywhere there are L1 learners could plausibly be, or is in fact, weaker in exoteric situations than a pressure *against* high E-complexity,

d. or how the predicted observations of such forces compare qualitatively or quantitatively (i.e. in relative strength) with neutral explanations of variation and change.

That is, without an explicit account of how high E-complexity ought to facilitate L1 learning it is difficult at minimum to understand its predictions or to evaluate it against empirical evidence. Second, without one or more neutral models of variation and change, there is nothing to compare either the empirical evidence or LNH predictions against, nor is it clear what the conditions are for hypothesized forces of selection to outweigh the effects of neutral forces — as opposed to being overwhelmed by them, as the next subsection notes is particularly plausible in small populations. Third, to really evaluate or understand the predictions of the LNH, we not only need to see an explicit model of L1 learning and its relation to E-complexity that supports the LNH, but also one of adult L2 learning and its relation to E-complexity. The reason why is that the posited causal mechanism behind the LNH's explanation of esoteric typology isn't actually something unique to esoteric situations — it's something present in *both* esoteric and exoteric situations (child learners) and a relative lack of something present in exoteric situations (adult L2 learners). As a result, any given variation and L1 learning model sufficient to predict selection *for* high E-complexity in all populations where there are child learners could end up predicting that the pressure for high E-complexity should in general prevail relative to any given adult L2 learning model sufficient to predict a preference by them *against* high E-complexity and for low E-complexity. Given the general expectation in evolutionary systems (elaborated

in §4.3.2) that drift is in general much stronger in small populations, and therefore only relatively strong forces of selection should be expected to reliably shape their evolution and so be a reasonable explanation for the typology of small populations, this concern is doubly important for the Linguistic Niche Hypothesis. Altogether this means that to be compatible with the full range of the LNH's predictions about E-complexity and social situation, any model of L1 learning offered in support of it that is sufficient to predict selection for high E-complexity in esoteric situations must also be weak enough relative to the selection pressure of a model of adult L2 learning sufficient to predict selection against high E-complexity in exoteric situations. Accordingly, understanding and evaluating the predictions of any model of L1 learning offered in support of the LNH's predictions about esoteric situations is partially dependent on what model of adult L2 learning is offered in support of the LNH's predictions about exoteric situations.

Turning to standards of evidence, Lupyan and Dale (2010, 2015, 2016b) and Dale and Lupyan (2011) spend little time considering or weighing neutral alternative explanations for the relationship between E-complexity and esoteric communities. Lupyan and Dale (2010, p. 8) does acknowledge drift briefly as an alternative hypothesis, but does not elaborate or discuss the relative strength of evidence for it. Dale and Lupyan (2011) contains no investigation or discussion of neutral mechanisms at all or what they would predict about either their agent-based simulation or their empirical investigation. Lupyan and Dale (2015) discusses 'drift', but does not use the term to describe a neutral random sampling-like process affecting which elements of a population survive and replicate,[20] but instead to describe two separate phenomena in an agent-based simulation of theirs. First, they use it to describe a linguistic analogue of *allopatric speciation*: when a population splinters into two or more geographically isolated populations, the populations may evolve along different evolutionary trajectories — 'drift apart', in this sense of Lupyan and Dale (2015)'s usage.

[20]I.e. what in the context of evolution 'drift' conventionally means.

This kind of divergence in evolution between geographically isolated populations is *not* synonymous with drift, the evolutionary force. Rather, it can be a consequence of neutral forces like drift, differing selection pressures in different environments (as in the simulation of Lupyan and Dale 2015), or some combination of both. In the case of Lupyan and Dale (2015)'s simulation the divergence in evolution of isolated groups is a consequence of a non-neutral migration model that preferentially keeps agents whose language variants are sufficiently similar together, geographically varying selection pressures, and the selective force their second usage of drift refers to. This second sense of 'drift' in Lupyan and Dale (2015) refers to an accommodation-like mechanism in their simulation whereby speakers adjust their linguistic representations to more closely match the average value in their local speech community — a frequency-dependent selection mechanism, not a neutral one. Finally, although a side-bar in Lupyan and Dale (2016b) correctly indicates that 'drift' in the context used in evolutionary theory refers to random sampling-like effects on which individuals survive and replicate, the main text only uses 'drift' to refer to the process and effects of a linguistic analogue of allopatric speciation. Neither Lupyan and Dale (2015) nor Lupyan and Dale (2016b), then, discuss or evaluate neutral explanations.

In the next two subsections, we discuss alternative explanations that do not require the assumption that there are different kinds of learners or that there is any selective pressure for high E-complexity specific to esoteric situations, but which still predict that small, esoteric populations should still be expected to display a greater degree of variation (§4.3.2) than large, exoteric ones, that even if a small population is subject to selection[21] drift is more likely to be the cause of evolutionary changes (§4.3.2), and that small populations are more likely to permit difficult-to-learn variants (if they exist) to persist or become common than in large, exoteric populations that are otherwise comparable (both §4.3.2 and §4.3.3).

[21]Regardless of whether it is specific to small populations.

4.3.2 Drift is a powerful force on small populations

Recall the basic structure of the LNH's explanation for the relationship between social situation and E-complexity presented in §4.2.3: the independent variables whose value or direction of change precedes all other steps in the causal chain of the LNH are demographic variables like population size. Among the kinds of forces – drift, migration, mutation, and selection – commonly examined in population genetics, drift is known[22] to be much stronger in small populations than large ones, and — for biologically plausible mutation rates and relative fitnesses — to be much more powerful than mutation or selection in small populations and negligible in large ones.

To illustrate this and its consequences for reasoning about what explains the observed state of an evolutionary system, consider again the hypothetical population discussed in §4.2.2 of a small population shown in Fig. 4.1. This is a possible evolutionary history of a population of 10 individuals with 2 possible trait types[23] over 5 generations in a variant of one of the basic models of population genetics: the Wright-Fisher model with drift, but no mutation, no migration, and no selection. Recall that this means that the generation at time t_{i+1} is created by sampling with replacement n times from the generation at time t_i, and this collection of samples constitutes the population at time t_{i+1}: the probability of any individual in the population at time t_i surviving and producing one replicant does not depend on the variant type of that individual and is uniform across the population.

Figure 4.2 illustrates what happens as we increase population size in this variant of the Wright-Fisher model: each graph shows the trajectories over 20 generations of 10 different populations that all start out with a 50/50 distribution over the two trait types. The

[22] See e.g. the population genetics textbooks Hartl and Clark (1997) or Rice (2004).

[23] To be more specific: in biological terms, these are individuals with one allele per gene (they are 'haploid') who reproduce asexually, and we are modeling the evolution of one locus ('gene') that can take on exactly one of two possible values ('alleles') and whose evolution is, by assumption, independent of all other loci in the organism's genome.

y-axis summarizes everything about the state of a population at a particular point in time in terms of the proportion of that population with one of the two trait values.[24] Figure 4.3 is the same, but for 1,000 generations of evolution. As population size gets larger, it's clear that drift has less and less effect per unit time: drift will take much longer, compared to when the population is small, to push a population's state a given distance from the same starting point.

These graphs illustrate several notable properties of drift as a force acting on a population:

(4.3.49) a. The absolute frequency of each variant undergoes fluctuations that are usually small at each step.

b. With one important exception (4.3.49c), fluctuation in one direction is as likely as any other — hence the name 'drift'. This is in contrast to other forces, like selection or potentially 'directed' neutral forces like asymmetric migration rates, as may e.g. be the case in a biological context between a a small island population and a larger mainland one.

c. Once a population evolving under drift contains no individuals of a variant type, that type will never appear again unless another source of variation (e.g. mutation or migration) re-introduces it.

Within a population, this means that in a small number of generations, drift causing a small number of changes in the absolute frequencies of a small population can cause a large change in relative frequencies: the population in Fig. 4.1 started evenly split over both variants, but it is quite likely that the population will end up consisting entirely of tokens of the 'blank' variant within a few timesteps of t_4. By the same token, the larger a population is, the less effect drift has on the trajectory of the population and the longer it will take for drift to cause

[24]What was 'blank' vs. 'filled' in Fig. 4.1 is here variant 'A' vs. 'a'.

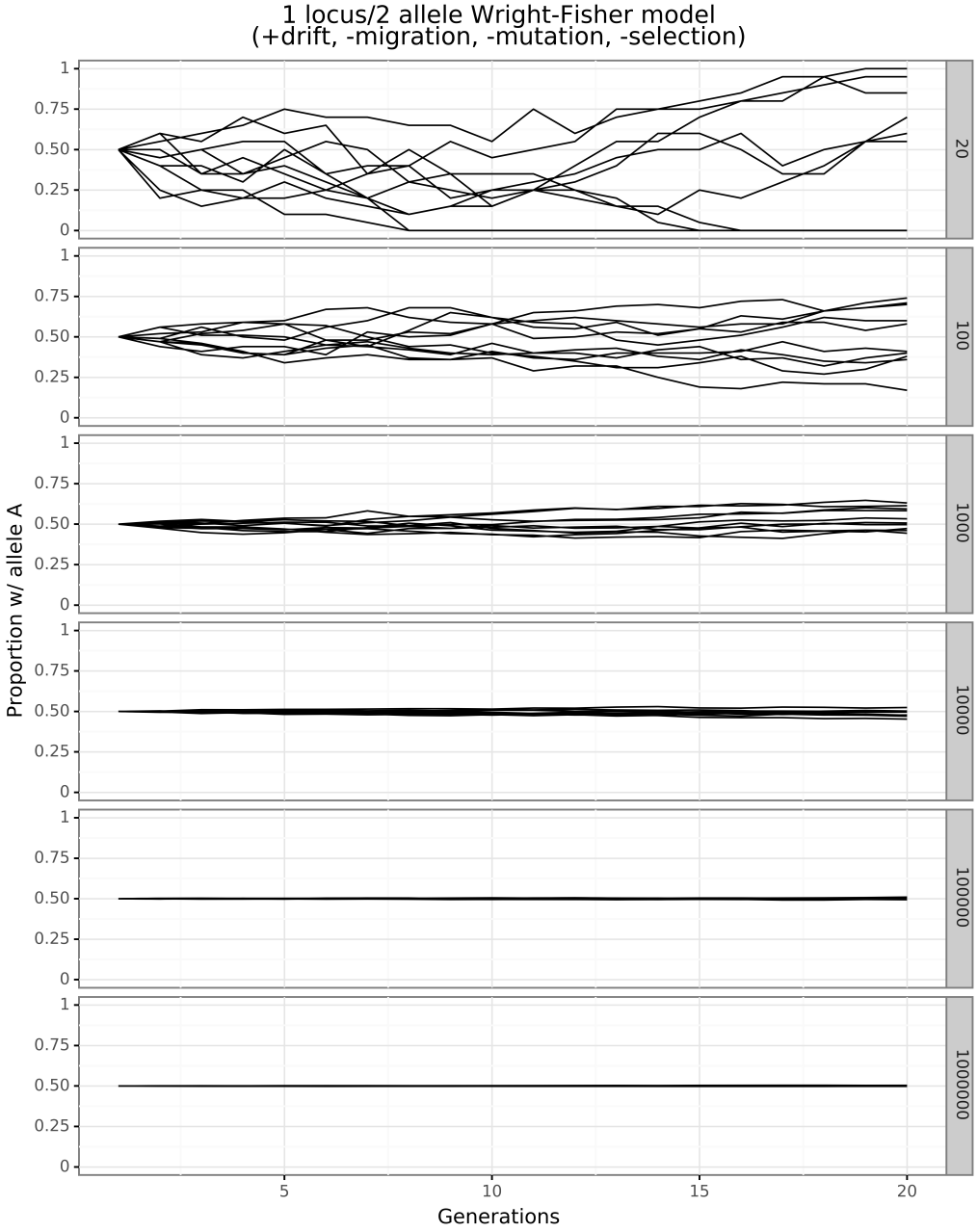


Figure 4.2: Each plot shows the trajectories (under drift alone) over 20 generations of 10 simulated populations with population sizes (indicated on the right) varying from 20 to 1,000,000.

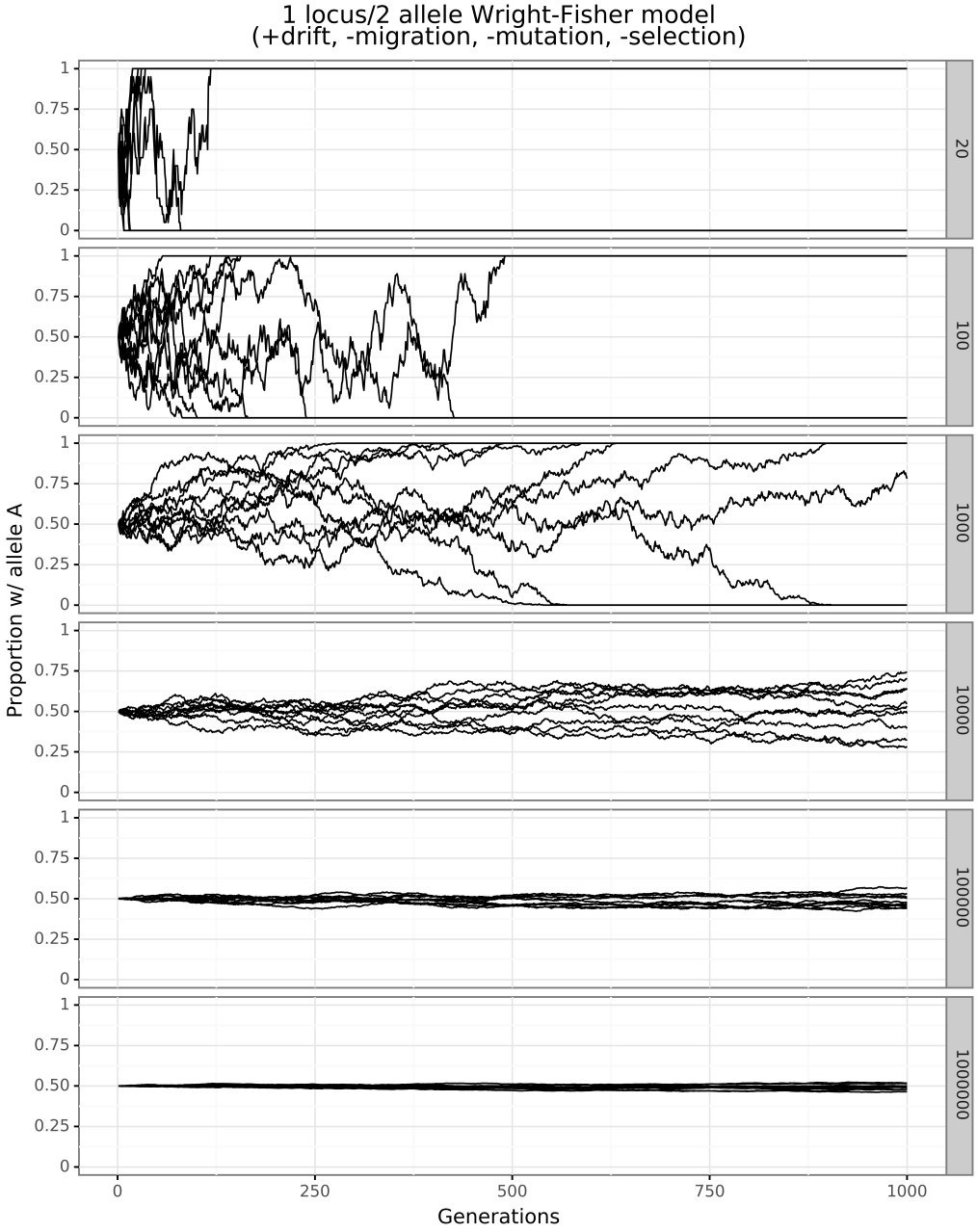


Figure 4.3: Each plot shows the trajectories (under drift alone) over 1,000 generations of 10 simulated populations with the population size (indicated on the right) varying from 20 to 1,000,000.

one variant vs. another to sweep to fixation. Drift also has important between-population effects: small subpopulations of the same species that are relatively separated from each other (due to e.g. geographic distance or other barriers) will each undergo drift, but do so separately (i.e. in uncorrelated directions). Without the intervention of other forces like high enough rates of migration or similar directed forces of mutation or selection operating in each subpopulation, the members of each subpopulation will likely become more similar to each other than to members of other subpopulations.[25] To summarize: with respect to a single population, the smaller the population, the stronger drift is as a source of long-term change — specifically loss of variation and increase in within-population homogeneity; with respect to multiple relatively separated and small subpopulations, drift is a force for diversification and divergence between those subpopulations.

In the context of morphology and esoteric populations, the takeaway is that, all else being equal, random fluctuations in replication frequency that are small in absolute number ought to be expected to have a much stronger effect on language change per unit time in a small community than in a large one, and that, all else being equal, drift will cause much more typological variation (including e.g. some amount of high E-complexity) across small populations on a given timescale than it will across large ones. Note that without any assumptions about selection for *or* against high E-complexity under any circumstances, drift alone should be expected to lead to more variation across a set of small populations at any given moment than between an otherwise comparable set of large ones.

How do the effects of drift *and* selection interact as a function of population size? Figures 4.4 and 4.5 are similar to the previous pair, except they now illustrate a Wright-Fisher model with a moderately strong amount of (frequency-independent) selection

[25]If this proceeds far enough for long enough, it can lead to allopatric speciation, referenced in the previous subsection.

for one of the two trait values.[26] As population size gets larger, the effect of drift becomes weaker and the direction and strength of selection becomes clearer.

With respect to the LNH, then, it is plausible that even if high morphological E-complexity *were* clearly and demonstrably advantageous for child learning relative to low morphological E-complexity, the effects of drift in small (esoteric) populations could plausibly mask or even overwhelm it. Generally speaking, it means that whatever forces of selection operate in all populations of speakers, drift should be plausibly expected to cause changes (typological variation) in relatively small populations that selection would be expected to filter out in relatively large ones. Reasoning about how likely this is as a relevant concern for the LNH, or what the general conditions are for this to likely be relevant[27] requires evidence about the relative strength of drift vs. different kinds of selection in language change — e.g. analysis of an explicitly presented model with mechanisms of drift and both L1 and adult L2 learning, more empirical data on learning, longitudinal data on population size, population structure, and E-complexity.[28]

Finally, these graphs should also drive home the importance of empirical data on how populations change over non-trival stretches of history for reasoning about what explains the typology we see currently. That is, consider the task of trying to determine the strength of evidence for selection in explaining the observed diversity of traits across several populations. Above we have simulated data for several such populations from a model that is an idealized, controlled, and oversimplified representation of biological evolution, and

[26]See any population genetics textbook for reference on the relevant calculations.

[27]I.e. for a given model of drift in language change, what counts as a large enough population size for drift to no longer have an appreciable effect on a particular timescale, in the absence of selection or in the presence of a particular kind and degree of selection? Given a particular model of drift, some choice of kind selection, and an empirically plausible degree of selection — whatever that may turn out to be — how large does a population need to be for the effects of selection to likely outweigh the effects of drift on a particular timescale?

[28]We refer the reader to a detailed exploration of several of the external factors of influence on language complexity in Bentz (2018).

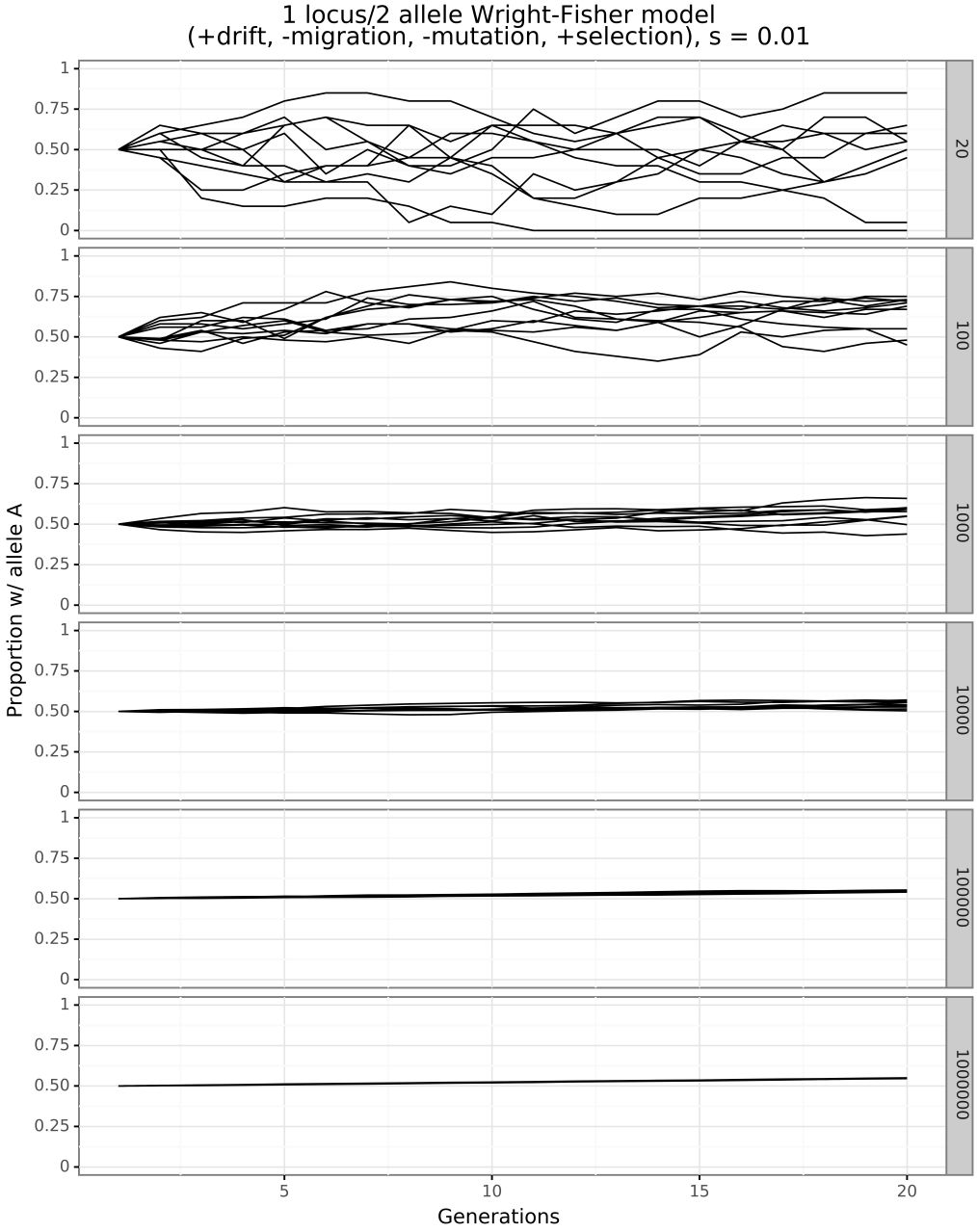


Figure 4.4: Each plot shows the trajectories (under drift and a moderate amount of selection) over 20 generations of 10 simulated populations with the population size (indicated on the right) varying from 20 to 1,000,000.

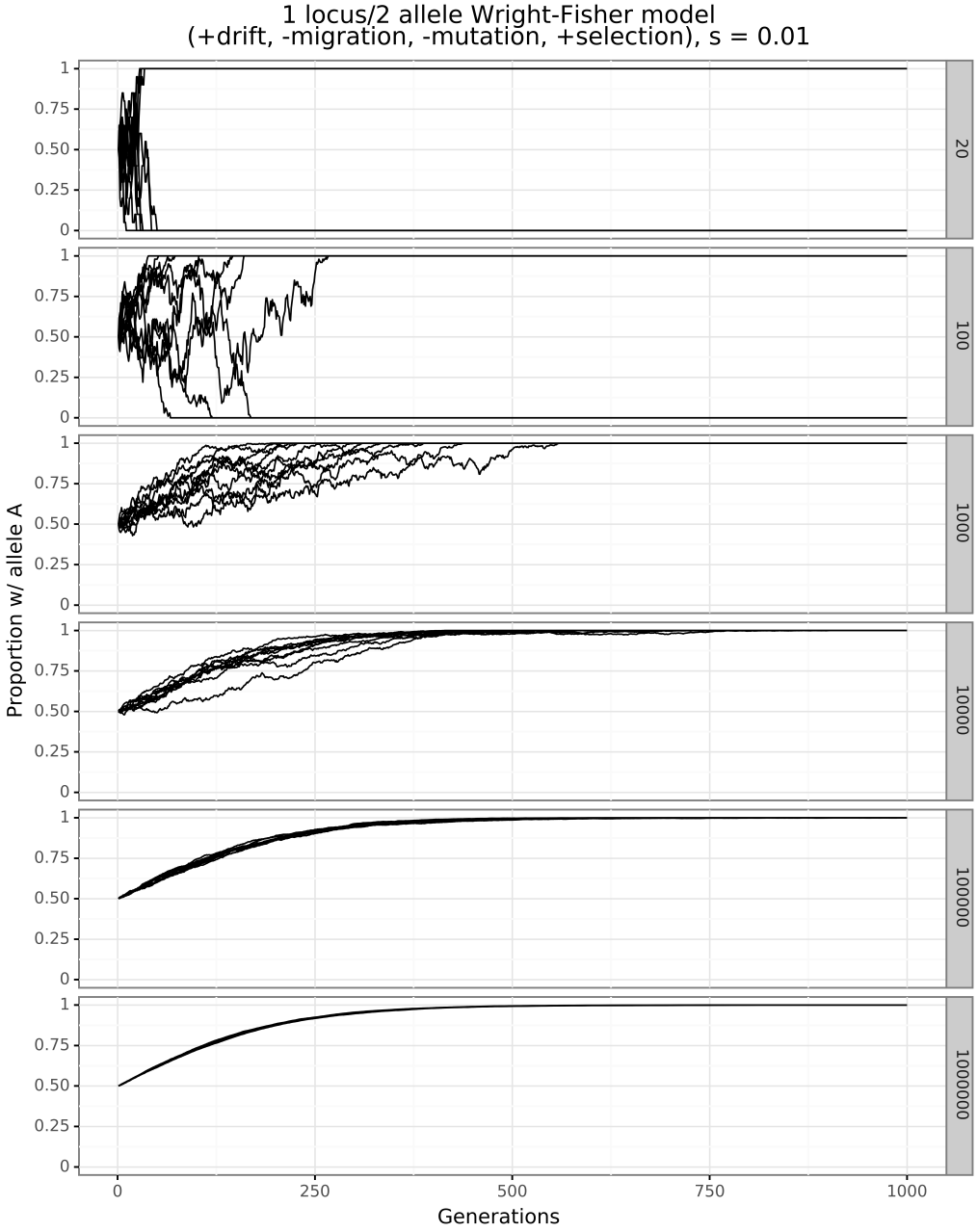


Figure 4.5: Each plot shows the trajectories (under drift and a moderate amount of selection) over 1,000 generations of 10 simulated populations with the population size (indicated on the right) varying from 20 to 1,000,000.

— crucially — we have many longitudinal measurements covering the entirety of a long timespan. In contrast, as noted earlier in §4.3.1, relatively little and sparse diachronic data is available about the linguistic structures, the relative 'fitness' of those structures, or the social structures for many of the languages and speech communities in Lupyan and Dale (2010)'s WALS-based analysis. Compare Figures 4.2 and 4.4, but imagine only seeing the state of a few populations from either graph and only seeing one or two points in time for each population. Under these conditions, determining with confidence whether a population is being acted on by selection or only subject to drift is extremely difficult, and our conclusions should be appropriately qualified and conservative (R. J. Smith, 2016).

In sum, the main independent variables behind the LNH's adaptive explanation of the relative prevalence of high E-complexity in esoteric situations should also be expected to amplify the effects of a much simpler neutral evolutionary force — drift. Drift is an 'undirected' force that should lead, all else being equal, to relatively large amounts of variation between small populations in a given span of time, and the relative strength of drift could plausibly overwhelm any effect of selection, if present. This also means that we should expect drift to cause relatively small populations to display linguistic variants that selection would remove in a larger population. Both drift alone and the LNH's two-part selection for high E-complexity in esoteric populations and selection against high E-complexity in exoteric ones are hypotheses that could predict observing more high E-complexity language variants in small populations. However, because drift is a strictly simpler explanation and more likely to explain differences in the evolutionary trajectories of small vs. large populations, drift should be regarded as an a priori more likely explanation than the LNH until we have clearer empirical evidence or model-based reasoning to suggest otherwise.

4.3.3 Relative homogeneity of input in esoteric populations

In this subsection we review two computational models of language change that manipulate the composition and structure of populations. Both indicate that, if there are difficult-to-learn linguistic variants (where difficulty is uniform across all learners), then small, esoteric populations are more likely to permit these difficult-to-learn variants to persist or become common than large, exoteric populations that are otherwise comparable, and that this is explainable as a consequence of differences in population size and structure in small, esoteric populations vs. large, exoteric ones rather than differences in which forces of selection are operating in esoteric vs. exoteric populations.

The first model is a variation on the Bayesian iterated language learning model outlined in §4.2.1. Whereas in the simplest form of this model each learner observes data produced by exactly one teacher who has chosen exactly one grammar as the basis for their productions, Burkett and Griffiths (2010, §4) and Dangerfield (2011) consider a more realistic setting where each learner's data comes from multiple teachers of the previous generation — and hence from multiple grammars. The task of learning is still reasoning about how likely different causes are to have given rise to the observed data, but now a 'cause' is a distribution over grammars rather than a single grammar. Accordingly, where the learners of Griffiths and Kalish (2007) discussed previously have a prior over grammars, learners in this multi-grammar setting have a prior over distributions of grammars. While a technical discussion of the form of this prior is outside the scope of this chapter, all that is important for the present discussion is that this prior has two parameters, a base distribution over grammars G_0 and a concentration parameter α. The base distribution is comparable to the prior over grammars discussed earlier, while the concentration parameter reflects the learner's expectations about both how many distinct grammars are responsible for the observed data and how close their distribution over grammars is to the base distribution:

$0 < \alpha < <1$ indicates an expectation that increasingly many datapoints are produced by very few grammars and where the distributions most likely to dominate are decreasingly close to the base distribution, while $1 < <\alpha$ indicates an expectation that increasingly many datapoints are produced by increasingly many grammars, and where the distribution over which grammars those are is increasingly close to the base distribution.[29] As the number of observations increases, the effect of a learner's prior diminishes and their posterior will approach the actual generating distribution; as α decreases, the rate at which this happens will increase. Given the relatively small number of observations per learner in Burkett and Griffiths (2010) and Dangerfield (2011), we are interested in moderate-to-lower values of α. In this parameter regime, the end result of iterated learning is an amplification of biases in the data presented to the initial population of learners.

That is, consider two instantiations of the model from Burkett and Griffiths (2010, specifically §4) or Dangerfield (2011, Ch. 5), one where the initial data are consistent with a relatively flat distribution over grammars — an *exoteric* starting condition with a relatively heterogeneous mix of grammars — and another where the initial data are consistent with a relatively peaked distribution over grammars — an *esoteric* starting condition with a relatively homogeneous mix of grammars. Absent some reason to expect an exoteric learner to observe more datapoints overall than an esoteric learner, a learner in the exoteric starting condition receives strictly fewer datapoints per language per unit time than an esoteric learner. This means that for any grammar variant g_{hard} that is more difficult to learn[30] than another grammar variant g_{easy}, exoteric learners will be less likely to end up selecting that grammar (given the same amount of data) than learners in a much more homogeneous population consisting principally of speakers who preferentially use g_{hard}. In the context of the Linguistic Niche Hypothesis, this means that if high E-complexity language variants are

[29] See Dangerfield (2011) for extensive discussion of the concentration parameter.

[30] I.e. require more observations on average for a learner to assign it a given probability.

indeed harder to learn for (all or any significant fraction of all) learners, then homogeneity of input in esoteric situations could be sufficient to allow harder-to-learn variants to be more likely to persist than in exoteric situations. Crucially, note that this does not require that high E-complexity be particularly *beneficial* to a type of learner that is specific to the esoteric social situation.

Reali et al. (2014) offer simulation results roughly mirroring the logic outlined above, but with three notable differences from Burkett and Griffiths (2010). First, where the model of Burkett and Griffiths (2010) is comparable to the discrete, non-overlapping generations Wright-Fisher model of population genetics where a replication event is synonymous with an abstract child language acquisition event, Reali et al. (2014) uses a model more comparable to the overlapping generations Moran model of population genetics where the replication process is comparable to individual episodes of production of a single utterance. Second, while the framework of Burkett and Griffiths (2010) does permit explicit manipulation and analysis of what variants require greater vs. fewer expected observations to acquire, Reali et al. (2014) do, in fact, explicitly manipulate the learning difficulty of linguistic variants. Third, where Burkett and Griffiths (2010) model every speaker-teacher from generation t as equally likely to contribute data for each new member of generation $t + 1$, Reali et al. (2014) assume a *spatialized* model where each speaker only interacts with nearby agents. While both Burkett and Griffiths (2010) and Reali et al. (2014) are neutral models insofar as the probability that any given speaker contributes data that influences a listener does not depend on their linguistic variant or distribution over variants, Reali et al. (2014) is both more realistic and specifically permits exploration of the idea that differences in the network structure of who talks to who in esoteric vs. exoteric communities contributes to differences in morphological typology (see e.g. Trudgill, 2009).

In more detail, Reali et al. introduce a kind of spatial structure to communicative interactions and allow the learnability of different linguistic conventions to vary. They

simulate a persistent population of communicating agents by placing each agent on a unique node in a type of random graph whose structure allows for gradient exploration of conditions corresponding to esoteric and exoteric social situations: as the number of nodes in the graph (population size) increases, the average number of neighboring nodes increases. Crucially, only agents in nodes that are connected ('neighbors') can communicate with each other. As a result, each agent in the esoteric condition tends to have repeated interactions with a small number of speakers who *themselves* tend to have repeated interactions with a small number of speakers (and so on…). Accordingly, a linguistic convention requiring relatively more observations to be accurately learned is more likely to perpetuate itself and take hold in an esoteric population than an exoteric one, all else being equal.

The models and results of both papers offer simpler alternative explanations of why small, esoteric populations are more likely to display variants that are harder for some portion of the population to learn than large, exoteric ones. Consider: the LNH's explanation depends on

(4.3.50) a. The existence of a force of selection for high E-complexity.

b. The hypothesis that this force is explained by a model of child learning favoring high E-complexity variants and leading to their preferential later use.

c. The hypothesis that this is specifically due to children having an easier time keeping track of redundant and explicit morphosyntactic information than reasoning about world knowledge or pragmatic information.

d. The force of selection from child learning for high E-complexity being strong enough relative to drift to influence the typology of esoteric populations.

e. The existence of a force of selection against high E-complexity.

f. The hypothesis that this force originates in adult L2 learning.

g. The force of selection for high E-complexity being weak enough relative to this

second force of selection against high E-complexity to explain the typology of exoteric situations.

In contrast, Burkett and Griffiths (2010) and Reali et al. (2014) suggest explanations that depend only on

(4.3.51) The existence of a force of selection against high E-complexity.

Note that with respect to both Burkett and Griffiths (2010) and Reali et al. (2014), this force of selection is rooted in learning preferences that are uniform over all agents in all populations. Both suggest that effects of population size and structure — effects that would also be present under the LNH's assumptions — can create conditions in small, esoteric populations that are plausibly sufficient to allow hard-to-learn variants to be maintained there at a higher rate than in large, exoteric ones. In other words, both papers suggest an explanation strictly simpler than the LNH.

4.4 Conclusion

We have argued that scientific explanations of variation and change in evolutionary systems (including language change) are beset by two key problems: a lack of informative data and a wealth of logically possible explanations with unclear or plausibly overlapping predictions. Further, we have argued that responsible scientific investigation in the face of these problems requires clearly presented and preferably mathematically explicit models of hypothesized mechanisms (like learning), thorough consideration of neutral explanations, and that simpler explanations be preferred by default. Turning specifically to the Linguistic Niche Hypothesis's adaptationist claim about the relationship between E-complexity and social situation, we have pointed out that both of the problems generally facing explanation in evolutionary systems are particularly acute for the Linguistic Niche Hypothesis, and that what mathematical models we do have suggest that there are simpler, neutral (or more

neutral) explanations for why high E-complexity (or generally, a language variant that is selected against in general) would be expected to be found in smaller, esoteric communities — i.e. explanations that do not invoke or require there to be any selection *for* high E-complexity specifically in esoteric social situations. First, we discussed how small population size should be expected to *amplify* the role of a neutral process (evolutionary drift) and *mask* the effects of selection in shaping the state and trajectory of an esoteric community's language variety. Second, we have reviewed recent work on mathematical modeling of language change suggesting that learnability selection *against* a language variant (crucially without selection *for* it in any condition) could lead to its differential appearance and persistence in small, esoteric populations by causing greater homogeneity of input to learners compared to exoteric situations.

In sum, we conclude that, in the absence of compelling evidence that high E-complexity facilitates child learning or the presentation of specific evidence against neutral explanations for the relation between morphological typology and social situation, general principles of evolutionary systems and current models of language change suggest that the most likely explanation for the morphological typology of esoteric communities does not reflect adaptation to infant learning. While the LNH was partly intended to account for supposed correlations between what we have denominated the E-complexity of morphological systems in esoteric situations, it, correctly, does not assume that languages in such situations are either the only languages with high E-complexity nor that they are necessarily more E-complex than those in exoteric situations. In fact, high E-complexity obtains for languages in very varied social situations, in many population sizes and ranging over different areal distributions. For example, Hungarian, a member of the Ob-Ugric branch of the Uralic language family with 13 million speakers, displays quite elaborate inventories of verbal and nominal marking, Mordvin, a member of the Volga-Finnic branch of the Uralic language family with approximately 400,000 speakers, possesses the most complex system of verbal

inflection in Uralic, while Navajo, a member of the Athapaskan family with approximately 145,000 speakers, contains an extraordinarily rich system of morphosyntactic and allomorphic variation in both its nominal and verbal systems (Bonami, McDonough, & Beniamine, 2019). From the perspective of parsimony, of course, we would like any account to cover the learning of all three languages, as well as languages exhibiting even more complex and simpler systems. Given this, a real learnability condundrum remains and becomes plain: how does the learning of (complex) morphological systems actually occur, in both small communities and larger ones, esoteric and exoteric? We have hypothesized that this process is guided by morphological organization measured in terms of I-complexity, i.e., patterns and subpatterns of conditional entropy that facilitate good guesses from known (patterns of) forms to unknown (patterns of) forms. Throughout we have alluded to connections between neutral theory, language change, and systemic morphological organization as synthesized in Lass (1997b).[31] While this is not the forum to develop these connections, we explore and explicate them in Ackerman, Bonami, and Malouf (2019).

Chapter 4 was coauthored with Farrell Ackerman and Robert Malouf, and is very similar to the submitted manuscript that has since been edited and will be published in the Cambridge University Press volume *Morphological Typology and Linguistic Cognition*. The dissertation author was the primary investigator and author of this paper.

[31] See also Norde and van de Velde (2016).

References

Ackerman, F., Blevins, J. P., & Malouf, R. (2009). Parts and wholes: Implicative patterns in inflectional paradigms. *Analogy in grammar: Form and acquisition*, 54–81.

Ackerman, F., Bonami, O., & Malouf, R. (2019). *Making Sense of Morphology*. Manuscript.

Ackerman, F., & Malouf, R. (2013). Morphological Organization: The Low Conditional Entropy Conjecture. *Language*, *89*, 429–464.

Ackerman, F., & Nikolaeva, I. (2014). *Descriptive typology and linguistic theory: A study in the morphosyntax of relative clauses*. Stanford: CSLI Publications.

Amundson, R. (1996). Historical development of the concept of adaptation. In M. Rose & G. V. Lauder (Eds.), *Adaptation* (pp. 11–53). Academic Press.

Amundson, R. (2005). *The changing role of the embryo in evolutionary thought: roots of evo-devo*. Cambridge University Press.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–517.

Anttila, R. (1989). *Historical and comparative linguistics*. Amsterdam: Benjamins.

Arthur, W. (2004). *Biased embryos and evolution*. Cambridge University Press.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, *47*(Pt 1), 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical*

Society of America, *119*(5 Pt 1), 3048–3058.

Bafumi, J., & Gelman, A. (2006). Fitting multilevel models when predictors and group effects correlate. *Available at SSRN 1010095*.

Baronchelli, A., Chater, N., Pastor-Satorras, R., & Christiansen, M. H. (2012). The Biological Origin of Linguistic Diversity. *PLoS ONE*, *7*(10).

Beddor, P. S. (2009). A Coarticulatory Path to Sound Change. *Language*, *85*(4), 785–821.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bentz, C. (2018). *Adaptive Languages: An Information-theoretic Account of Linguistic Diversity (Vol. 316)*. Walter de Gruyter GmbH & Co KG.

Blevins, J. (2004). *Evolutionary Phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Blevins, J. (2008). Natural and Unnatural Sound Patterns: A Pocket Field Guide. In K. Willems & L. De Cuypere (Eds.), *Naturalness and iconicity in language* (pp. 121–148).

Blevins, J. (2016). *Word and Paradigm Morphology*. Oxford University Press.

Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, *88*(2), 269–304.

Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives* (Unpublished doctoral dissertation). University of Amsterdam.

Bonami, O., & Beniamine, S. (2016). Joint predictiveness in inflectional paradigms. *The Behavior Analyst*, *9.2*, 156–182.

Bonami, O., & Henri, F. (2010). Assessing empirically the inflectional complexity of Mauritian Creole. In *Workshop on formal aspects of creole studies.* Berlin.

Bonami, O., McDonough, J., & Beniamine, S. (2019). When segmentation helps: Implicative structure and morph boundaries in the Navajo verb. *Submitted*, *Manuscript*.

Bonami, O., & Strnadova, J. (2018). Paradigm structure and predictability in derivational morphology. *Morphology*, *Online Preview*, 1–31.

Bradley, T. G. (2001). A typology of rhotic duration contrast and neutralization. *Proceedings of the North East Linguistic Society*, *31*, 79–98.

Burkett, D., & Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8), Utrecht, Netherlands, 14-17 April 2010*, 58–65.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016a). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*(July), 68–86.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016b). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers ' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86.

Bybee, J. (1985). *Morphology: A Study of the Relation Between Meaning and Form*.

Bybee, J. (2001). *Phonology and Language Use*.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. (2010). The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. In *Language resources and evaluation* (Vol. 44, pp. 387–419).

Carbonell, K. M., & Lotto, A. J. (2014). Speech is not special...again. *Frontiers in Psychology*, *5*.

Chang, S., Plauche, M., & Ohala, J. J. (2001). Markedness and consonant confusion asymmetries. In *The role of speech perception in phonology* (chap. 4).

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Chater, N., Reali, F., & Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences*, *106*(4), 1015–1020.

Chomsky, N. (1957). *Syntactic Structures*.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher corpus: a Resource for the Next Generations of Speech-to-Text. *Language Resources and Evaluation*, *4*, 69–71.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–253.

Clayards, M. A. (2008). *The Ideal Listener: Making optimal use of acoustic phonetic cues for word recognition* (Unpublished doctoral dissertation). University of Rochester.

Cohen Priva, U. (2008). Using information content to predict phone deletion. *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 90–98.

Cohen Priva, U. (2012). *Sign and Signal Deriving Linguistic Generalizations From Information Utility* (Doctoral dissertation). Stanford University.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Corning, P. (2018). *Synergistic Selection: How Cooperation Has Shaped Evolution and the Rise of Humankind*. World Scientific.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory.* John Wiley & Sons.

Croft, W. (2000). *Explaining Language Change: An Evolutionary Approach*.

Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, *10*(3), 261–273.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668–3678.

Dale, R., & Lupyan, G. (2011). Understanding the origins of morphological diversity: the linguistic niche hypothesis. *Advances in Complex Systems*.

Dale, R., & Lupyan, G. (2012). Understanding The Origins Of Morphological Diversity:

The Linguistic Niche Hypothesis. *Advances in Complex Systems*.

Dangerfield, K. (2011). *Iterated learning of language distributions* (MSc). Edinburgh.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128–145.

Davies, M. (n.d.). *The Corpus of Contemporary American English (COCA).*

Degen, J. (2013). *Alternatives in Pragmatic Reasoning* (Unpublished doctoral dissertation). University of Rochester.

Demberg, V., Sayeed, A. B., Gorinski, P. J., & Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*(July), 356–367.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Ehala, M. (1996). Self-Organisation and Language Change. *Diachronica*, *13*, 1–28.

Elman, J. L., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective On Development*. The MIT Press.

Embick, D. (2015). *The morpheme: A theoretical introduction (Vol. 31)*. Walter de Gruyter GmbH & Co KG.

Esper, E. A. (1925). A technique for the experiment investigation of associative interference in artificial linguistic material. *Language monographs*.

Esper, E. A. (1966). Social transmission of an artificial language. *Language*, *42*, 575–580.

Feldman, N. H., & Griffiths, T. L. (2009). A Rational Account of the Perceptual Magnet Effect. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 1–6.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Flemming, E. (2001a). *Auditory representations in phonology*.

Flemming, E. (2001b). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, *18*, 7–44.

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162–178.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Gallagher, G. (2012). Perceptual similarity in non-local laryngeal restrictions. *Lingua*, *122*(2), 112–124.

Garrett, A. (2015). Sound change. In C. Bowern & B. Evans (Eds.), *The Routledge Handbook of Historical Linguistics* (pp. 227–248).

Garrett, A., & Johnson, K. (2011). Phonetic bias in sound change. *UC Berkely Phonology Lab Annual Report*, *2008*, 9–61.

Geisler, W. S. (2003). Ideal Observer Analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825–838). Cambridge: MIT Press.

Godfrey, J. J., & Holliman, E. (1997). *Switchboard-1 Release 2* (Tech. Rep.). Linguistic Data Consortium.

Gould, S., & Lewontin, R. R. C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Program. *Proceedings of the Royal Society B: Biological Sciences*, *205*(1161), 581–598.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics (Vol. 1)*. New York: Wiley.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive science*, *31*(3), 441–480.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology.* Cambridge University Press.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & psychophysics*, *28*(4), 267–283.

Hale, J. T. (2003). *Grammar, Uncertainty, and Sentence Processing* (Unpublished doctoral dissertation). The Johns Hopkins University.

Hale, J. T. (2006). Uncertainty About the Rest of the Sentence. *Cognitive Science*, *30*, 643–672.

Hale, M., & Reiss, C. (2000). "Substance Abuse" and "Dysfunctionalism": Current Trends in Phonology. *Linguistic Inquiry*, *31*(1), 157–169.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (2018). The Role of Predictability in Shaping Phonological Patterns. *Linguistic Vanguard*, *4*.

Hansson, G. Ó. (2008). Diachronic Explanations of Sound Patterns. *Language and Linguistics Compass*, *2*(5), 859–893.

Hartl, D. L., & Clark, A. G. (1997). *Principles of Population Genetics* (3rd ed.). Sunderland, MA: Singauer Associates.

Hayes, B. (1999). Phonetically driven phonology: The role of Optimality Theory and inductive grounding. *Functionalism and Formalism in Linguistics, Volume 1: General Papers*, 243–285.

Hayes, B., & Steriade, D. (2004). Introduction: the phonetic bases of phonological markedness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 1–33). Cambridge University Press.

Hayes, B., & White, J. (2013). Saltation and the P-map. *Phonology*, 1–23.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation*(2009), 187–197.

Hood, K. E., Halpern, C. T., Greenberg, G., & Lerner, R. M. (Eds.). (2010). *Handbook of developmental science, behavior, and genetics*. Wiley.

Hull, D. L., & Ruse, M. (Eds.). (2008). *The Cambridge Companion to the Philosophy of Biology*. Cambridge University Press.

Hume, E., & Johnson, K. (2001). A Model of the Interplay of Speech Perception and Phonology 1. In *The role of speech perception in phonology* (chap. 1).

Hura, S. L., Lindblom, B., & Diehl, R. L. (1992). On the role of perception in shaping phonological assimilation rules. *Language and speech*, *35*, 59–72.

Iggesen, O. A. (2013). Number of cases. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Ince, R. A. A. (2017). Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy*, *19*(318).

Irvine, L., Roberts, S. G., & Kirby, S. (2013). A robustness approach to theory building: A case study of language evolution. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2614–2619.

Jablonka, E., & Lamb, M. J. (2014). *Four dimensions of evolution: Genetic, epigenetic, behavioral and symbolic variation in the history of life*. Cambridge MA: MIT Press.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*.

Jaeger, T. F., & Buz, E. (2017). Signal Reduction and Linguistic Encoding. In E. Fernandez & H. Cairns (Eds.), *Handbook of psycholinguistics.* Wiley-Blackwell.

Jaeger, T. F., & Buz, E. (2018). Signal Reduction and Linguistic Encoding. In *The handbook of psycholinguistics* (pp. 38–81). Wiley-Blackwell.

Janda, L., & Tyers, F. M. (2018). Less is more: why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory*, *Online Preview*, 1–33.

Jun, J. (1995). *Perceptual and Articulatory Factors in Place Assimilation: An Optimality Theoretic Approach* (Unpublished doctoral dissertation). UCLA.

Jun, J. (2004). Place assimilation. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 58–86). Cambridge University Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. *Frequency and the emergence of linguistic structure*, 229–254.

Kauhanen, H. (2017). Neutral change. *Journal of Linguistics*, *53*(2), 327–358.

Keating, P. A. (1985). Universal Phonetics and the Organization of Grammars. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (chap. 8).

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Blackwell.

Kirby, J. (2013). The role of probabilistic enhancement in phonologization. *Origins of sound change: Approaches to phonologization*.

Kirby, S. (2001). Spontaneous evolution of linguistic structure - An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Kohler, K. J. (1990). Segmental Reduction in Connected Speech in German: Phonological Facts and Phonetic Explanations. In *Speech production and speech modelling* (pp. 69–92).

Kruszewski, M. (1995). Outline of linguistic science. In E. Koerner (Ed.), *Writings in general linguistics: "on sound alternation" (1881) and "outline of linguistic science" (1883)*. Amsterdam: John Benjamins.

Kusters, W. (2003). *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*.

Laland, K. (2018). *Darwin's unfinished symphony: How culture made the human mind*. Princeton University Press.

Lass, R. (1997a). Explanation and ontology. In *Historical linguistics and language change* (chap. 7).

Lass, R. (1997b). *Historical linguistics and language change*. Cambridge University Press.

Levy, R. (2005). *Probabilistic Models of Word Order and Syntactic Discontinuity* (Doctoral

dissertation). Stanford University.

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 234–243).

Lewontin, R. C. (1970). The Units of Selection. *Annual Review of Ecology and Systematics*, *1*.

Lewontin, R. C. (1978). Adaptation. *Scientific American*, *239*(3), 212–230.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439).

Lindblom, B., Guion, S., Hura, S., Moon, S.-J., & Willerman, R. (1995). Is sound change adaptive? *Rivista di Linguistica*, *7*, 5–37.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*, 3–11.

Łubowicz, A. (2003). *Contrast preservation in phonological mappings* (Unpublished doctoral dissertation). University of Massachusetts Amherst.

Łubowicz, A. (2007). Paradigmatic Contrast in Polish. *Journal of Slavic Linguistics*, *15*(2), 229–262.

Luce, P. a. (1987). Neighborhoods of Words in the Mental Lexicon. *Dissertation Abstracts International, B: Sciences and Engineering*, *47*(12).

Lupyan, G., & Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE*, *5*.

Lupyan, G., & Dale, R. (2015). The role of adaptation in understanding linguistic diversity. In R. D. Busser & R. J. LaPolla (Eds.), *The shaping of language* (chap. 11).

Lupyan, G., & Dale, R. (2016a). Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. *Trends in Cognitive Sciences*, *20*(9), 649–660.

Lupyan, G., & Dale, R. (2016b). Why are there different languages? The role of adaptation

in linguistic diversity. *Trends in Cognitive Sciences*.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Maiden, M. (2018). *The Romance verb*. Oxford University Press.

Malécot, A. (1956). Acoustic Cues for Nasal Consonants: An Experimental Study Involving a Tape-Splicing Technique. *Language*, *32*(2), 274–284.

Malécot, A. (1958). The Role of Releases in the Identification of Released Final Stops: A Series of Tape-Cutting Experiments. *Language*, *34*(3), 370–380.

Marr, D. (1982). The Philosophy and the Approach. In *Vision: A computational investigation into the human representation and processing of visual information* (pp. 8–38). San Francisco: W.H. Freeman and Company.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522–523.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science*, *189*(4198), 226–228.

Marslen-Wilson, W., & Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1–71.

Marslen-Wilson, W., & Tyler, L. (1987). Against modularity. In *Modularity in knowledge representation and natural language understanding* (pp. 37–62).

Marslen-Wilson, W., Tyler, L., & Seidenberg, M. (1976). Sentence processing and the clause boundary. In *Studies in the perception of language* (pp. 219–246).

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

McClelland, J. L., & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, *18*.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*(1), 509.

McShea, D. W., & Brandon, R. N. (2010). *Biology's First Law*.

Meakins, F., Hua, X., Algy, C., & Bromham, L. (2019). Birth of a contact language did not favour simplification. *Language*, 1–44.

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of The Royal Society*, *363*(September), 3489–3501.

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*(2), 338–352.

Moore, D. S. (2006). The developmental systems approach and the analysis of behavior. *The Behavior Analyst*, *39.2*, 243–258.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*, 83–127.

Moreton, E., & Pater, J. (2012a). Structure and Substance in Artificial-Phonology Learning. *Linguistics and Language Compass*, *6*(11), 702–718.

Moreton, E., & Pater, J. (2012b). Structure and Substance in Artificial-Phonology Learning, Part II: Substance. *Linguistics and Language Compass*, *6*(11), 702–718.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178.

Newell, A. (1981). The Knowledge Level: Presidential Address. *AI Magazine*, *2*(2), 1.

Newell, A., & Simon, H. A. (1961). *Computer simulation of human thinking*.

Norde, M., & van de Velde, F. (2016). *Exaptation and language change*. Amsterdam: John Benjamins.

Norris, D. (1994). Shortlist - a Connectionist Model of Continuous Speech Recognition. *Cognition*, *52*(3), 189–234.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Norris, D., & Kinoshita, S. (2012). Reading through a noisy channel: Why there's nothing special about the perception of orthography. *Psychological Review*, *119*(3), 517–545.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Norris, D., Mcqueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18.

Ohala, J. (1993). *The phonetics of sound change.*

Ohala, J. J. (1974). Experimental historical phonology. In J. M. Anderson & C. Jones (Eds.), *Proceedings of the 1st international conference on historical linguistics. edinburgh, 2-7 sept. 1973.*

Ohala, J. J. (1975). Phonetic Explanations for Nasal Sound Patterns. *Nasálfest: Papers from a symposium on nasals and nasalization*, 289–316.

Ohala, J. J. (1981). The Listener as a Source of Sound Change. In *Papers from the parasesson on language and behavior: Chicago linguistics society* (pp. 178–203).

Ohala, J. J. (1990a). The phonetics and phonology of aspects of assimilation. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology i: Between the grammar and physics of speech* (pp. 258–282).

Ohala, J. J. (1990b). *There is no interface between phonology and phonetics: a personal view* (Vol. 18).

Ohala, J. J. (1992). What's cognitive, what's not, in sound change. In *Diachrony within synchrony: language history and cognition* (pp. 309–355).

Ohala, J. J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, *13*, 155–161.

Ohala, J. J. (1997). Phonetics in Phonology. In *Proceedings of the 4th Seoul international conference on linguistics [SICOL] 11-15 aug 1997* (pp. 45–50).

Oyama, S., Gray, R. D., & Griffiths, P. E. (2001). *Cycles of contingency: Developmental systems theory and evolution*. Cambridge, MA: MIT Press.

Padgett, J. (1997). Perceptual Distance of Contrast: Vowel Height and Nasality. , *5*(1957), 63–78.

Padgett, J. (2003). The Emergence Of Contrastive Palatalization In Russian. In *Optimality theory and language change* (pp. 307–335).

Padgett, J. (2009). Systemic contrast and Catalan rhotics. *The Linguistic Review*, *26*(4), 431–463.

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*.

Pater, J., Jesney, K., & Smith, B. (2012). Learning probabilities over underlying representations. In *SIGMORPHON '12 Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology* (pp. 62–71).

Perkins, R. D. (1992). *Deixis, Grammar, and Culture*.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pigliucci, M., & Kaplan, J. (2000). The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later. *TREE*, *15*(2), 66–70.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye corpus of conversational speech (2nd release).*

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*.

Ramscar, M., Dye, M., Blevins, J., & Bayyaen, H. (2018). Morphological development. In A. Bar-On & D. Ravid (Eds.), *Handbook of communication disorders theoretical, empirical, and applied linguistic perspectives* (pp. 181–202). De Gruyter.

Reali, F., Chater, N., & Christiansen, M. (2014). The paradox of linguistic complexity and community size. In *The evolution of language: Proceedings of the 10th international conference* (pp. 270–277).

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society B*, *277*,

429–436.

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: lexical stress drives eye movements immediately. *Quarterly journal of experimental psychology (2006)*, *63*(4), 772–783.

Reiss, C. (2017). Substance Free Phonology. In S. J. Hannahs & A. R. K. Bosch (Eds.), *Handbook of phonological theory.* Routledge.

Rice, S. H. (2004). *Evolutionary Theory: Mathematical and conceptual foundations*.

Riedl, R. (1977). A Systems-Analytical Approach to Macro-Evolutionary Phenomena. *The Quarterly Review of Biology*, *52.4*, 351–370.

Robinson, J. A. (1965). A machine-oriented logic based on the resolution principle. *Journal of the ACM*, *12*, 23–41.

Rosenberg, A., & McShea, D. W. (2008). *Philosophy of Biology: A contemporary introduction*.

Rubino, C. (2013). Reduplication. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Sadat, J., Martin, C. D., Costa, A., & Alario, F. X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive Psychology*, *68*, 33–58.

Sampson, R. (1999). *Nasal vowel evolution in Romance*.

Sarkar, S., & Plutynski, A. (Eds.). (2008). *A Companion to the Philosophy of Biology*. Blackwell.

Scholz, B. C., Pelletier, F. J., & Pullum, G. K. (2011). *Philosophy of Linguistics.* Retrieved from `https://plato.stanford.edu/entries/linguistics/`

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex systems*, *1*, 145–168.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Sims, A. D. (2015). *Inflectional defectiveness*. Cambridge: Cambridge University Press.

Sims, A. D., & Parker, J. (2016). How inflection class systems work: On the informativity of implicative structure. *Word Structure*, *9.2*, 215–239.

Sims, A. D., & Parker, J. (2019). Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of Russian nouns. In P. Arkadiev & F. Gardani (Eds.), *Morphological complexity* (pp. 1–35). Manuscript.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Smith, R. J. (2016). Explanations for Adaptations, Just-So Stories, and Limitations on Evidence in Evolutionary Biology. *Evolutionary Anthropology*, *25*, 276–287.

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: a database of Dutch diphone perception. *The Journal of the Acoustical Society of America*, *113*(January), 563–574.

Stephens, C. (2008). Population Genetics. In S. Sarkar & A. Plutynski (Eds.), *A companion to the philosophy of biology* (pp. 119–137). Blackwell.

Steriade, D. (2001a). Directional asymmetries in place assimilation: a perceptual account. In E. Hume & K. Johnson (Eds.), *Perception in phonology.* Academic Press.

Steriade, D. (2001b). *The Phonology of Perceptibility Effects: The P-Map and Its Consequences for Constraint Organization.*

Steriade, D. (2008). The Phonology of Perceptibility Effects: The P-Map and Its Consequences for Constraint Organization. In K. Hanson & S. Inkelas (Eds.), *The nature of the word* (pp. 150–178). MIT Press.

Stolcke, A. (2002). SRILM-An Extensible Language Modeling Toolkit. In *8th international conference on spoken language processing (interspeech 2002)* (Vol. 2, pp. 901–904).

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proceedings - IEEE automatic speech recognition and understanding workshop.*

Stone, J. V. (2015). *Information theory: a tutorial introduction*. Sebtel Press.

Stump, G., & Finkel, R. A. (2013). *Morphological Typology*. Cambridge University Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, *268*, 5–10.

Thurston, W. (1987). *Processes of change in the languages of north-western New Britain*.

Thurston, W. (1992). Sociolinguistic typology and other factors effecting change in north-western New Britain, Papua New Guinea. In T. Dutton (Ed.), *Culture change, language change; case studies from melanesia* (pp. 123–139).

Trudgill, P. (2009). Sociolinguistic typology and complexification. In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language complexity as an evolving variable* (chap. 7).

Trudgill, P. (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*.

Trudgill, P. (2016). The sociolinguistics of non-equicomplexity. In R. Baechler & G. Seiler (Eds.), *Complexity, isolation, and variation.*

Turnbull, R., Seyfarth, S., Hume, E., & Jaeger, T. F. (2018). Nasal place assimilation trades off inferrability of both target and trigger words. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1).

Tyler, L., & Marslen-Wilson, W. (1977). The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 683–692.

Van Son, R. J. J., Koopmans-van Beinum, F. J., & Pols, L. C. W. (1998). Efficiency As An Organizing Principle Of Natural Speech. In *Fifth international conference on spoken language processing*.

Van Son, R. J. J., & Pols, L. C. W. (2003). How efficient is speech? In *Proceedings of the institute of phonetic sciences* (pp. 171–184).

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, *28*(4), 735–747.

Vitevitch, M. S., & Luce, P. A. (2016). Phonological Neighborhood Effects in Spoken Word Perception and Production. *Annual Review of Linguistics*, *2*, 75–94.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: a study of perceptual features. *The Journal of the Acoustical Society of America*, *54*(5), 1248–1266.

Warner, N., McQueen, J. M., & Cutler, A. (2014). Tracking perception of the sounds of English. *The Journal of the Acoustical Society of America*, *135*(5), 2995–3006.

Warner, N., Smits, R., McQueen, J. M., & Cutler, A. (2005). Phonological and statistical effects on timing of speech perception: Insights from a database of Dutch diphone perception. *Speech Communication*, *46*(1), 53–72.

Wayland, S. C., Wingfield, A., & Goodglass, H. (1989). Recognition of Isolated Words: The Dynamics of Cohort Reduction. *Applied Psycholinguistics*, *10*(4), 475–487.

Wedel, A., & Fatkullin, I. (2017). Category competition as a driver of category contrast. *Journal of Language Evolution*, *2.1*, 77–93.

White, J. (2014). Evidence for a Learning Bias Against Saltatory Phonological Alternations. *Cognition*, *130*(1), 96–115.

White, J. (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language*.

Whyte, L., Lancelot. (1965). *Internal factors of evolution*. Tavistock Publications.

Wilson, C. (2003). Experimental Investigation of Phonological Naturalness. *West Coast Conference on Formal Linguistics 22 (WCCFL22)*, 101–114.

Wilson, C. (2006). Learning Phonology With Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science*, *30*(5), 945–982.

Winter, B., & Wedel, A. (2016). The Co-evolution of Speech and the Lexicon: The Interaction of Functional Pressures, Redundancy, and Category Variation. *Topics in Cognitive Science*, *8.2*, 503–513.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers : Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*, 543–578.

Wright, R. (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (chap. 2). Cambridge University

Press.

Wurzel, W. (1987). System-dependent morphological naturalness in inflection. In W. Dressler (Ed.), *Leitmotifs in natural morphology* (pp. 59–96).

Zhang, J. (2001). *The Effects of Duration and Sonority on Contour Tone Distribution – Typological Survey and Formal Analysis* (Unpublished doctoral dissertation). UCLA.

Zhang, J., & Lai, Y. (2006). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Kansas Working Papers in Linguistics*, *28*, 65–126.

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*, 153–201.

Zipf, G. K. (1936). *The Psychobiology of Language*. London: Routledge.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning corpus and survey evidence from Tagalog infixation. *Language*, *83*, 277–316.

Zuraw, K. (2013). **MAP Constraints.* Retrieved from `http://www.linguistics.ucla.edu/people/zuraw/dnldpprs/star_map.pdf`