To appear in Language and Linguistics.

Probabilistic phonology: A review of theoretical perspectives, applications, and problems

John Alderete and Sara Finley
Simon Fraser University/Pacific Lutheran University

Probability and frequency are becoming increasingly important in phonological analysis. This article reviews contemporary perspectives on how phonological theory addresses gradient phonological patterns shaped by probability and frequency, drawing on theories of the lexicon, grammar, and statistics. After examining their motivations, we show how these diverse theoretical perspectives have been applied to a variety of problems in core phonology, including phonotactics, morphophonology, sound change, phonological categorization, and language development. Our review of theory and applications supports a growing consensus in the field that phonological theories must reckon with probability. Our review also identifies problems stemming from a lack of cohesion in the field, and suggests potential solutions to these problems.

Keywords: probability, frequency, gradience, exemplar theory, information theory, constraint-based grammar, connectionism

1. Introduction

Throughout much of its history, generative phonology has been focused on explaining categorically-defined phenomena using a binary classification of phonological form (i.e., grammatical vs. ungrammatical). While early work on the scope of generative grammar did entertain degrees of grammaticality (Chomsky 1961; Chomsky and Miller 1963; Katz 1964), and quantified phonological variation in non-categorical terms (Labov 1969), it is fair to say that the challenge for most phonologists has been to account for the distinction between grammatical and ungrammatical forms. Marginal or graded phenomena that fall between the two extremes have tended to be addressed through the distinction between competence and performance (Chomsky 1957; Chomsky 1965; Chomsky and Halle 1968). In this view, phonological patterning on the margins is part of performance, and so it is not part of phonological analysis and can be disregarded as "noise".

However, over the last several decades, the field has shown a sizable shift of focus from the categorical to the gradient, and this shift is clearly evident in generative phonology. It has become increasingly clear that many gradient phenomena cannot be relegated to performance, and are indeed a part of linguistic competence (for review, see Ernestus (2011), Shaw and Kawahara (2018), Pierrehumbert (forthcoming)). To take a prominent and historically important example, dissimilatory co-occurrence constraints (as in OCP place constraints in Arabic) require reference to probabilistically-determined conditioning factors, such as locality, target features, and featural combinations (Pierrehumbert 1993; Frisch et al. 2004; Coetzee and Pater 2008; Hayes and Wilson 2008). Gradient phonotactic constraints such as these are reflected in native speaker judgments of both known words and nonsense words (Frisch et al. 2000; Bailey and Hahn 2001; Frisch and Zawaydeh 2001; Myers and Tsay 2005; Daland et al. 2011), and are therefore likely to be a part of the speaker's grammatical knowledge, rather than artifacts of historical residue or random variation. These developments in assessing phonological well-formedness mirror similar ones in syntactic analysis (Keller 2000; Sorace and Keller 2005), suggesting that linguistic competence includes gradient representations across a variety of domains.

Furthermore, a growing number of studies have shown that the application of phonological processes is correlated with a host of gradient phonological structures that are missed under a categorical approach to well-formedness. Phonological processes are influenced by lexical statistics (Ernestus and Baayen 2003; Hayes and Londe 2006; Zuraw 2007), social factors (Labov 1969; Cedergren and Sankoff 1974), and information-carrying capacity (Cohen Priva 2015; Hall et al. 2018; Shaw and Kawahara 2018), in addition to a range of phonological structures and contexts. In sum, a phonological system cannot be accurately characterized with a generative model that is restricted to categorical outcomes. Rather, non-categorical outcomes must also be entertained to describe the full range of phonological behavior.

The need to characterize phonology in terms of the gradient and the probabilistic has led to a number of innovations in theoretical models. This growth in approaches to probability in phonology suggests a growing consensus that phonological grammars must assign a role for frequency and probability. This role has

been formalized in almost every known analytical framework, including modifications to traditional generative models (e.g., Labov 1969), psycho-linguistically inspired models like exemplar theory (Pierrehumbert 2003b; Wedel 2006), and analogical models of phonology (Bybee 2001). In addition, mathematically-informed models such as information theoretic (Hume 2008; Hall 2009; Cohen Priva 2012), connectionist (Hare 1990; Alderete and Tupper 2018), and constraint-based models of phonology (Zuraw 2000; Goldwater and Johnson 2003; Hayes and Wilson 2008) have also been developed. Though there is considerable diversity in theoretical approaches, there exists an over-arching notion of probabilistic phonology that unifies these different perspectives. Empirically, probabilistic phonology involves embracing probability distributions for describing phonological structures and processes, above and beyond simple binary characterizations of grammaticality. Theoretically, we can say that probabilistic phonology is an enterprise that directly employs probability distributions in a phonological analysis, or assigns a role to a continuous measure that correlates with probabilities.

This article provides a brief background on probability theory that serves as the backbone to these theories (section 2), as well as an overview of these theoretical perspectives (section 3) and their applications in phonological analysis (section 4). This article is not intended to be a detailed argument for any particular theoretical position, or as an argument that a specific phonological domain requires reference to probabilities (though we do review distinct perspectives and illustrate differences among them). We also do not intend to give complete reviews of particular theoretical models. The reader is invited to review the detailed arguments given in the works cited in this paper, including several articles that focus on specific theories: Shaw and Kawahara (2018) on information theoretic phonology, Wedel (2006) and Ernestus and Baayen (2011) on exemplar models, Alderete and Tupper (2018) on connectionist phonology, Daland (2014) and Chandlee and Heinz (2016) on computational phonology, and Coetzee and Pater (2011) on Harmonic Grammar and variation.

Because progress has been made in understanding the role of probability in phonology from so many diverse perspectives, it can be a challenge to understand how each approach fits in to the larger goal, and how to compare approaches that may appear divergent, despite similar goals. This paper works to fill this gap. Our principal goal here is to try to develop a broader perspective and explicate a shared commitment to grapple with gradient patterns that spans across a range of frameworks. The theoretical perspectives we review, though they share a commitment to probability theory, differ considerably in underlying assumptions and the specific methods they employ. With these differences, come problems in contrasting and comparing accounts of the same phenomena, and these difficulties lead to a kind of insularity across models that unfortunately prevents cross-communication of ideas. The second contribution of this article is to lay out some of these problems in comparative analysis and also foreground some productive strategies for finding common ground within probabilistic phonology (section 5).

2. Background

2.1 Gradience, frequency, and probability

It is fairly common for phonological patterns to be described as 'gradient' in the sense that the correct analysis of a phonological pattern requires reference to non-integer numbers. In variable phonology (see Table 1a), for example, we can speak of a 0.7 probability of rule application, rather than categorically applying 100% or 0% of the time (with probabilities 1 or 0). Likewise, incomplete neutralization and related phenomena (Table 1b) are sometimes described as gradient in a similar sense because they result in fine-grained phonetic structures that require continuous variables. For example, in the neutralization case, vowel durations can be increased by 10% (or the duration can be scaled by 1.1) before devoiced obstruents that are underlyingly voiced (Dinnsen and Charles-Luce 1984).

While such uses of the term gradient are fairly common, we are also concerned with a more literal sense of this term in which sub-classes of a phonological pattern fall on a gradient. In this sense, familiar from linear regression, values on a continuous dimension are correlated with other measures and the correlation line may have a slope. For example, phonological forms exhibit gradient acceptability on a well-formedness scale, and the values of forms on this scale have been shown to be correlated with predictor variables, which are themselves continuous (Table 1c). To illustrate, Hayes and Wilson (2008) show that children's rating data for nonce onset clusters are positively correlated with MaxEnt values (i.e., transformations of harmony values, or weighted sums of constraint violations). Phonological classes falling on a gradient have also been used to

account for phonological classes that emerge from lexical statistics (Table 1d) and phonological grades in allomorphy (Table 1e).

Table 1. Empirical overview

a. **Variable phonological processes:** Processes that do not apply categorically, but instead whose likelihood of application is conditioned by a range of factors (e.g., structure description, social factors, speech rate). Example: in Panamanian Spanish, the likelihood of spirantization of the syllable-final alveolar flap is conditioned by word position, morpheme class, segment natural class, and socio-economic status (Cedergren 1973; Cedergren and Sankoff 1974)
b. **Incomplete neutralization:** Fine-grained phonetic structure of the output of a phonological process that is unexpected given its phonological categorization. Example: small but statistically significant differences between lexical and derived voiceless obstruents (for review, see Warner et al. (2004)).
c. **Gradient acceptability of word forms:** Rating of grammatical well-formedness, correlated with a continuous dimension. Example: the acceptability of unattested English onsets (e.g., *vz, ml, fr*) is strongly correlated with the MaxEnt values representing the combined impact of weighted constraints on licit onsets (Hayes and Wilson 2008).
d. **Gradient patterns in lexical statistics:** Phonological patterns that fall on a scale in how well they are represented in the lexicon. Example: the frequency of consonant pairings in the same root is scaled by the similarity of the two consonants in Arabic, as in, *s-ʃ* < *s-t* < *s-l* (Frisch et al. 2004).
e. **Graded morpho-phonology:** Allomorphy in which related forms fall into grades. Example: allomorphs for the English past tense have grades between strongly irregular (*sing/sung*), regular *(walk/walked*), and intermediate categories with the regular form (*feel/felt, cut*) that children overgeneralize (Bybee and McClelland 2005).

How does phonological theory reckon with probabilities other than 0 and 1, and how does it predict these gradients? In some models, probability and frequency are used to predict phonological behavior. In information-theoretical phonology, for example, a number of predictor variables derive from contextual probabilities, and so probability is encoded directly in the analysis, as in the use of informativity in Cohen Priva (2015). Other theories do not use probability or frequency directly, but instead use theoretical constructs that approximate probabilities or track frequencies. In MaxEnt Grammar, for example, well-formedness constraints are assigned weights that maximize the probability of an observed form, and these weights are generally arrived at through learning systems working with large data sets (Hayes and Wilson 2008). The weights themselves are not probabilities, or frequencies, but they have the effect of maximizing the probability of attested forms when couched within grammar. As we shall see, the use of weights that are interpreted as probabilities, or correlated with them, is quite common in constraint-based models, connectionist phonology, and exemplar phonology, perhaps because of the ubiquity of the use of weights in mainstream connectionist psycholinguistic models (Goldrick 2007).

2.2 Synopsis of probability theory

We give a brief synopsis of probability theory in order to explain how concepts from this theory are employed by the theories discussed below, but see Bod (2003) and Frisch (2012) for more detailed introductions to probability and frequency in linguistics.

Table 2. Definitions.

a. Probability of A in sample space $\Omega = \#A/\#\Omega$, if all outcomes equally likely
b. Joint probability $P(A, B) = P(A) \times P(B)$, if A, B independent

c. Conditional probability $P(B\|A) = P(A, B)/P(A)$, if $P(A) \neq 0$
d. Joint probability $P(A, B) = P(A) \times P(B\|A)$, if A, B dependent
e. Transitional probability $P(A \rightarrow B) = \#AB/\#A$
f. Contextual predictability $P(x\|\text{Context}) = \#x$ in Context/#Context
g. Surprisal of x in Context = $-\log_2 P(x\|\text{Context})$
h. Entropy of Context = $-\Sigma_x P(x\|\text{Context}) \log_2 P(x\|\text{Context})$
i. Informativity of $x = -\Sigma_c P(\text{Context}\|x) \log_2 P(x\|\text{Context})$
j. Bayes' theorem: $P(\text{A}\|\text{B}) = [P(\text{B}\|\text{A}) \times P(\text{A})] / P(\text{B})$

Probabilities can be viewed either as facts about the world that are calculated from the outcomes of experiments (frequentist interpretation) or as degrees of belief by an observer (objectivist or Bayesian interpretation). In both interpretations, probabilities are numbers between 0 and 1, where 0 indicates impossibility and 1 certainty. For example, a fair coin will have a probability 0.5 of landing Heads. Because most of the methods reviewed in this paper are based in frequentist assumptions, except where explicitly noted otherwise, the examples in this section also take a frequentist approach.

On this view, we calculate probabilities with experiments that produce observable outcomes. The collection of outcomes resulting from an experiment is called the sample space Ω, and any subset of Ω is an event. If all possible outcomes of an experiment are equally likely, then we can say that the probability of an event A, or $P(A)$, is a ratio of the size of A and the size of Ω (see

Table 2 for formulas). For example, suppose we have a corpus of 100 words in which 90 words begin with a consonant and the rest begin with a vowel. We can devise an experiment that selects a word from this corpus at random. What is the probability that a word will begin with a consonant on a trial of this experiment? The sample space is the entire 100 word corpus, and the event we are interested in is the subset of 90 words that begin with a consonant, so $P(\text{W}_{\text{Consonant}}) = 90/100 = 0.9$.

The probability of a linguistic structure is often estimated with frequency within a corpus (token frequency) or a dictionary/lexicon (type/lexical frequency). To illustrate the difference, the phonemes [*v*] and [*ð*] have relatively high token frequencies in English because they occur in many high frequency items, like *the* and *of*. However, these sounds have much lower type frequency because they are used in only a small number of words. That is, there are relatively few words other than *the* and *of*. Thus, the probability of hearing a given consonant depends on the way words and segments are counted, as well as the specific frequency distribution of the segments, as probability distributions over segments are not evenly distributed (Daland 2013).

Much of the interest in using probability distributions to study phonology comes from studying combinations of structures and structures in context. In an experiment where two structures A and B are sampled independently, and the selection of one does not affect the other, the chance of observing A and B together, referred to as the joint probability of A and B, is simply the product of $P(A)$ and $P(B)$. However, in many phonological studies, we may wish to investigate two distinct outcomes that are not independent. To calculate joint probability in such a scenario, we require another term, conditional probability $P(B|A)$, or the probability of B given A. The general definition for $P(B|A)$, is $P(A, B)$ over $P(A)$, though $P(B|A)$ can also be estimated from the data itself. The joint probability of A and B (i.e., $P(A, B)$, when A and B are dependent), is the product of $P(A)$ and $P(B|A)$.

To make these terms concrete, consider again a hypothetical corpus of 100 CV words in which 60 begin with an obstruent and 50 contain front vowels. We can devise an experiment that samples consonants from the words of this corpus at random, and samples vowels from different random words, and then replaces the words back to the corpus after sampling from it. In such an experiment, the sampling of consonants and vowels are independent, and so the joint probability of sampling an obstruent and a front vowel is simply

$P(C_{obstr}) \times P(V_{front})$, or $0.6 \times 0.5 = 0.3$. Suppose, however, we devise a different experiment that also selects words from the corpus at random, but first samples the consonant, then the vowel, of the same word. In such an experiment, the two outcomes are not independent, and so we need to know the conditional probability of obstruents given front vowels in order to calculate their joint probability. It turns out that on closer inspection obstruents are associated with front vowels because they occur in 45 of 50 words that have front vowels in the corpus. From these facts, we know the conditional probability of an obstruent occurring given a front vowel to be $45/50 = 0.9$. The joint probability of an obstruent occurring with front vowels in this experiment is thus the probability of obstruents ($60/100 = 0.6$) times the probability of an obstruent occurring given a front vowel, $P(C_{obstr}|V_{front}) = 0.9$. Thus, the joint probability $P(C_{obstr}, V_{front})$ is $P(C_{obstr}) \times P(C_{obstr}|V_{front}) = 0.6 \times 0.9 = 0.54$.

There are other ways in which context can be formalized in probabilistic phonology. One common approach is with transitional probability, the probability of a transition from structures *A* to *B*. Transitional probability is calculated as the frequency of the bigram *AB* over the frequency of *A*, and is often used in accounts of speech segmentation and finite-state accounts of phonology (Heinz 2010). Transitional probabilities have also been used in discussions of speech segmentation (Brent and Cartwright 1996; Saffran et al. 1996), where the transitional probability is higher within a word than at a word boundary. In the phrase *pretty baby*, the probability that [prɪ] comes before [ti] is higher than the probability that [ti] comes before [beɪ], because *pretty* is a word, and [ti]-[beɪ] is a transition between words (Saffran 2003).

Another way in which context can be incorporated probabilistically is through formalisms from information theory (Shannon and Weaver 1949). Information theory provides a set of measures for quantifying the information carrying potential of a message (see Cohen Priva (2015) and Shaw and Kawahara (2018) for review). An important concept in this theory is surprisal, which can be thought of as the number of bits of information a structure *x* holds if no other information aside from the context is known. Surprisal is calculated by taking the negative log of the contextual predictability of some structure *x* relative to a specified Context, $P(x|\text{Context})$. Other information theoretic measures are entropy, the average predictability of structure in a given context, and informativity, a measure of the amount of information a structure usually has in the language as a whole (see Table 2 for formulas). Entropy and informativity thus build on surprisal to give more general characterizations of the information-carrying capacity of linguistic structure.

While the above discussions focused on frequentist assumptions, many applications of conditional probabilities in both frequentist and Bayesian approaches make use of Bayes' rule. Suppose, for example, an observer hears a CV syllable from the corpus described above, but is only able to make out that the consonant is an obstruent, and would like to make an inference about the quality of the vowel (i.e., front versus back). Using the known information about the corpus, an inference about the probability of a front vowel given an obstruent, $P(V_{front}|C_{obstr})$, can be obtained. According to Bayes' rule, an unknown conditional probability $P(V_{front}|C_{obstr})$ (referred to as the posterior) is equal to the known conditional $P(C_{obstr}|V_{front})$ (referred to as the likelihood) multiplied by the individual probability of the hypothesis $P(V_{front})$ (referred to as the prior), divided by the evidence (or marginal likelihood), $P(C_{obsr})$. Using the formula in Table 2j, $P(V_{front}|C_{obstr}) = [P(C_{obstr}|V_{front}) \times P(V_{front})] / P(C_{obstr}) = [0.9 \times 0.5] / 0.6 = 0.75$.

In many of the examples discussed here, probability is calculated based on a finite set of forms, based on a corpus or results from a behavioral experiment. However, an important goal of generative phonology is to predict which forms are possible given a grammar, and not simply the existing words. Consistent with the latter goal, generative frameworks sometimes produce an infinite set of possible words from a finite set of constraints, rules, or representations. Defining a probability distribution over an infinite set may be a challenge for probabilistic models of generative phonology. Daland (2015) shows that while MaxEnt models of Harmonic Grammar (Hayes and Wilson 2008) appear to be especially prone to this issue, a straightforward solution is to incorporate constraints against excess structure (*Structure constraints) into the model. Other work in Optimality Theory (OT) has shown that while OT produces an infinite candidate set, there is generally only a finite set of possible winning candidates (contenders), thus reducing the probability space (Eisner 2002; Riggle 2004). While the problem of infinite sets has been discussed to some degree for OT and Harmonic Grammar, particularly in terms of the candidate set, it is not clear that this issue has been addressed at length for other frameworks, or for defining the probability distribution of all possible words. This is an issue for future research to address.

3. Theoretical perspectives

The motivation to formalize frequency and probability distributions with concrete theoretical mechanisms arises from a variety of different theoretical perspectives. We review some of the most prominent perspectives below, and also establish the role of probability and related notions in these theories.

3.1 Variable Rules

One of the earliest approaches to assume that phonology is probabilistic in nature is the Variable Rules framework (Labov 1969; Cedergren and Sankoff 1974; Guy 1991; Guy 1992). A fundamental assumption in this framework is that phonological processes have variant realizations that cannot be straightforwardly characterized with phonological grammars that require deterministic outcomes. Instead, variable phonology must be analyzed with statistical models that bring out the probabilistic aspects of both linguistic and non-linguistic factors. In its classic implementation (Labov 1969), variable phonology is analyzed as standard re-write rules, where the application of a process is expressed in terms of a probability. Statistical modeling is used to analyze the impact of conditioning factors in the structural description of the re-write rule (the independent variables) on the probability of a rule's application (the dependent variable). After some experimentation with the appropriate statistics, consensus arose in the variationist literature that logistic regression is the most appropriate analysis for this kind of modeling because of certain problems with linear probability models (see Paolillo (2002) for review). Many analyses make use of VARBRUL software, developed originally in the 1970's, with multiple updates, including GoldVarb (Sankoff et al. 2005). This software package provides tools for managing input data, constructing contingency tables, and conducting the logistic regression analysis (see Paolillo 2002, Appendix 1 for access to the software in different languages, software support, and related statistical packages). The regression analysis produces an input probability, which is the overall probability that the rule will apply, as well as the probability weights for conditioning factors.

To illustrate the Variable Rules approach with an example, Panamanian Spanish spirantizes alveolar flaps to an apical fricative in a variable pattern that is affected by position within a word, morphology, following segments, and socio-economic status (Cedergren and Sankoff 1974). Paolillo (2002: 33-34) gives an analysis of this variable pattern using logistic regression. In his analysis, spirantization has an input probability of 0.333, but contextual factors can increase or reduce its probability. For example, the existence of a following lateral, which has a probability weight of 0.743, leads to a higher probability of spirantization, 0.561. Thus, the inclusion of the lateral results in a change from a roughly 1-in-3 chance (basic input probability) to a 1-in-2 chance. As this example illustrates, probability is critical to the analysis of variable patterns: the overall analysis is stated as a probability distribution, and the individual factors impacting the larger outcome are expressed as probability weights (though whether analyses produced by VARBRUL itself induce probability distributions over the entire data set, or instead introduce weights that we interpret as probabilities, is a matter of debate; see (Eisner 2002)). Variable Rules phonology was one of the first fully developed theories linking phonological grammar, in the form of SPE rules, to probability.

3.2 Information theory phonology

Much research in probabilistic phonology has built upon insights from information theory (Shannon and Weaver 1949; Pierce 1961) to incorporate these insights into phonological analysis (Aylett and Turk 2004; Hume 2008; Hall 2009; Cohen Priva 2012; Turnbull 2015; Daland and Zuraw 2018; Hashimoto 2021). Like many approaches to probabilistic phonology, information theory typically analyzes data from large corpora, producing counts of structures of interest and the contexts they occur in. The corpus data are then used to calculate the surprisal, entropy, and informativity of these structures relative to a specified context (see section 2 for definitions and formulas). These measures can either be employed directly to give quantitative assessment of a phonological pattern, like the degree to which two structures contrast (Hall 2012), or they can be employed as predictor variables in larger statistical analyses. One advantage of information theoretic analyses of phonology is that they have natural connections to the psycholinguistics of speech perception and production (Hall 2009; Turnbull 2015), which creates the potential for linguistic discovery and new insights into classic problems (see Shaw and Kawahara (2018) and Pierrehumbert (forthcoming) for review).

One case for information theory in phonology is that it can take context into account across the entire language, rather than just the context for the phonological pattern in question. For example, prior research has shown that predictability within a context can affect segment and syllable durations: highly predictable

structures have shorter durations and can lead to segment deletion (Aylett and Turk 2004; Pluymaekers et al. 2005). However, contextual predictability is not the only factor, because structures with low contextual predictability can reduce and delete, as in *d* in the word *sudden*, and structures with high predictability can also be preserved. Cohen Priva (2015) argues that cases like this can be addressed if one considers a role for informativity (see section 2), or the amount of information a structure usually has across all contexts in a language, rather than in a specific context. In particular, this study used a set of regression models to examine the impact of segment probability, informativity, and contextual probability on consonant duration and incidence of deletion in English. It found that higher informativity leads to longer durations and a lower incidence of deletion, even when these other factors are controlled for. This approach, and others like it, demonstrate the importance of probability in phonology: surprisal and informativity derive directly from contextual probabilities, and they are important predictors of phonological processes like reduction and segment deletion.

3.3 Phonology with Bayesian inference

Bayes' rule (defined in Section 2) makes use of known or prior information to make inferences about unknown data, allowing the observer to make decisions among competing alternatives. For this reason, Bayes' rule is especially applicable in any case where there is ambiguity in the signal, such as spoken language processing (Norris and McQueen 2008), spoken language production (Kirov and Wilson 2013), and second language processing (Wilson and Davidson 2013). Bayes' rule has been widely implemented in a variety of learning algorithms related to probabilistic phonology, including learning phonological rules (Goldwater and Johnson 2004; Goldsmith and Riggle 2012) and phonological categories (Feldman et al. 2009), in addition to word learning (Frank et al. 2007; Xu and Tenenbaum 2007) and speech segmentation (Norris and McQueen 2008; Goldwater et al. 2009; Daland and Pierrehumbert 2011).

While the mathematical formulation of Bayes' rule is relatively simple, there are a variety of choices that need to be made related to determining prior probabilities, the hypothesis space, the sampling method, and how the model and hypothesis space is updated as the learner obtains more observations of the data. Bayesian models therefore have a wide range of applications, but each application may have differing assumptions. For example, Wilson and Davidson's (2013) Bayesian account of second language processing integrates phonotactic probability with perceptual likelihood, but the predictions of the model vary depending on how phonotactic probabilities are defined.

3.4 Exemplar phonology

Another way in which phonological representations have been argued to show gradient effects is in terms of frequencies in the lexicon. In exemplar models of the lexicon, word representations are stored as the collective of individual memories of utterances, and the overall representation is the average of all exemplar representations (Pierrehumbert 2001b; Pierrehumbert 2003a; Wedel 2003). Speakers encode all individual utterances of a lexical item, both spoken and perceived, thereby keeping track of statistical information such as frequency, context, variability, and fine-grained phonetic details. This kind of representation can directly accommodate many of the factors at work in probabilistic phonology, such as frequency, salience, and talker-specific representations. One potential issue with exemplar models is that if every utterance is stored, including ungrammatical utterances and misperceptions, over time, everything should become a possible word. To address this, Wedel (2003) proposes that representations in exemplar models have weights that track frequency, relevance, etc., so that more frequent and more appropriate productions have higher weights, while the weights of infrequent or misperceived tokens may decay over time (see Tupper (2015) for mathematical analysis of this approach within a field model). The strengths of different cues can change over time, depending on factors such as frequency and confusability.

Exemplar phonology is compatible with a theory of phonology in which there are many levels of abstraction in phonological representations, that is, continuous phonetic space, discrete phonological categories, word-forms, etc. (Pierrehumbert 2003b; Beckman and Edwards 2010). One proposal for such a system is that abstract, categorical phonological rules (such as vowel harmony) emerge out of highly detailed lexical representations within the lexicon (Bybee 2001; Pierrehumbert 2001a). Another possibility is that exemplar models contain a function that allows speakers to generalize to novel items (Nosofsky 1986). For example, the probability that a speaker will confuse two different segments relates to their phonetic similarity, which in turn correlates with the probability of a sound change (Bybee 2001; Blevins 2006; Johnson 2006;

Garrett and Johnson 2013). A similarity function could also be used to determine the probability with which a novel lexical item is likely to be inferred as grammatical by a speaker of the language. For example, abstract processes like vowel harmony can be accounted for in exemplar phonology, so long as the similarity is calculated based on properties of the vowels, and the morphological composition of the words in question (Cole 2009).

3.5 Analogical models of phonology

Exemplar models can also account for generalization through analogy. In analogical models of phonological representations, generalization to novel items occurs through analogy to known items (Skousen 1989; Skousen 1992; Skousen 1995; Eddington 2000; Bybee 2001), or through the spreading of activation among words that have similar form and meaning (Daelemans et al. 2002). A lexical item selected in analogy is dependent on three factors: proximity (e.g., similarity), gang effects (e.g., the number of items that behave similarly), and heterogeneity (e.g., the lack of more probable alternatives). Analogical models that analyze these factors are therefore compatible with usage-based approaches to linguistic structure (Bybee 2001), and can be used to make predictions about the probability of a given phonological form. For example, an analogical model that assigns stress patterns as a function of these three factors (Eddington 2000) has been shown to successfully predict stress placement in 94% of Spanish words. Analogical models have also been used to demonstrate how phonological variables predict morpho-phonological operations, such as the English past tense (Eddington 2004) and stem alternations in Dutch (Ernestus and Baayen 2003).

3.6 Constraint-based models

While usage-based models of language, such as exemplar and analogical models, have been relatively vocal about gradience in phonological processes, generative models have also recently begun to assign a role for gradience and probability. In particular, constraint-based models have established a role for grammar in the analysis of probabilistic phonology. In these models, grammars are constraint systems that optimize over candidates for an input-output mapping. In OT (Prince and Smolensky 1993/2004), variation between languages is accounted for with the interaction of a universal set of constraints; differences between languages are a result of different constraint rankings. The problem of learning a grammar is one of finding the constraint ranking for a given language, and OT is readily used as a complement to learning theories (Tesar and Smolensky 2000; Prince and Tesar 2004; Jarosz 2006). Classic OT (Prince and Smolensky 1993/2004) assumes a single, categorical outcome. However, there are several variations of OT, including models with built-in learning algorithms, that have been successfully adapted to account for probabilistic patterns.

For example, Boersma and Hayes' (2001) Stochastic OT model is an early example of using OT to account for both learning and variability. Learning involves gradually adjusting the ranking value of constraints on a continuous numerical scale, which establishes the ranking relationships among constraints required for determining grammatical outcomes. Variable outcomes arise in this model with an evaluation procedure that, for each constraint, permutes the ranking values by drawing on a normal distribution centered on its learned value. By introducing this noise in the model, constraints with overlapping distributions produce variability in the output.

Other constraint-based approaches make use of Harmonic Grammar (HG), a co-development of OT, in which constraints are weighted rather than strictly ranked (Legendre et al. 1990; Smolensky and Legendre 2006). In Harmonic Grammar, candidate outputs are assessed with a so-called harmony score, calculated as the weighted sum of a given candidate's constraint violations. The candidate with the highest harmony score is selected as the output. Like in Stochastic OT (Boersma & Hayes 2001), noise can also be introduced in constraint evaluation and used to account for phonological variation (Goldrick and Daland 2009; Coetzee and Pater 2011). Importantly, variation arises from the grammatical architecture, because it is produced by the noisy evaluation of constraints, the stuff of grammars. As discussed in section 5, because each candidate is assigned a harmony score, Harmonic Grammar can approximate gradience in lexical statistics through phonological learning (Coetzee and Pater 2008).

Harmony has also been employed in Maximum Entropy (MaxEnt) models where harmony scores for a candidate set are converted to probability scores, which indicate the probability of producing any given output form (Goldwater and Johnson 2003; Hayes and Wilson 2008); see also Daland (2015) for a formal analysis. Another use of harmony, Gradient Symbol Processing, combines an optimization dynamics similar to Harmonic Grammar with a quantization dynamics that pulls outputs toward discrete symbolic categories

(Smolensky et al. 2014). This approach has been successful at modeling gradient patterns in speech errors (Goldrick and Blumstein 2006). Finally, constraint-based grammars can also make use of analogy (section 3.5), whereby paradigmatic operations across words support generalizations within the lexicon and guide learning (Becker and Gouskova 2016).

3.7 Connectionist phonology

As noted above, Harmonic Grammar uses a similar architecture to 'classic' OT, except that Harmonic Grammar makes use of weighted, or soft constraints, rather than strict rankings. Another way to implement soft constraints to account for probabilistic phonology is through activation dynamics in connectionist networks (Goldsmith 1993; Alderete and Tupper 2018). A connectionist network is a web of interconnected micro-processors or "units". Soft constraints are analyzed as connections between these micro-processors (Smolensky 1988), and are satisfied if the receiving unit resembles the state of the sending unit (for positive connections) or resembles the opposite state (for negative connections). The strength of the constraint is the connection's weight (or the weights of many connections), parallel to Harmonic Grammar (Smolensky and Legendre 2006).

Like OT and Harmonic Grammar, connectionist models are learning models. The weights of connections start at random and are learned through exposure to a large data set. After training, connectionist networks produce outcomes that can, like Harmonic Grammars, approximate probabilistic patterns very well (St. John and McClelland 1988; Thomas and McClelland 2008). Connectionist models have been successful at accounting for gradience, variability, and sensitivity to frequency. For example, Laks' (1995) connectionist model of French syllabification (based on ideas from Goldsmith and Larson (1990)) parallels the graded intuitions that French speakers have of different syllabic roles.

4. Applications

This section explores how the theoretical frameworks discussed above can be applied to a variety of gradient phenomena, including more accurate characterizations of phonological processes, phonological variation, sound change, and language learning. While these processes have implications for psycholinguistic processing of phonology, we do not specifically review applications to probabilistic phonology in psycholinguistics here (but see Jurafsky (2003)).

4.1 Linguistic discovery and description

The theories discussed in section 3 make use of computational methods and typically work with large data sets (corpora). The availability of a range of computational toolkits for exploring phonology in corpora (Bird and Loper 2004; Rose et al. 2006; Myers 2012; Durand et al. 2014; Hall et al. 2016) has led to new discoveries and insights into known phenomena. To illustrate with a simple example, Bird et al. (2004: 103) show how to use the Natural Language Toolkit to uncover phonological generalizations in Rotokas, a Northern Bougainville language of Papua New Guinea. With just a few lines of Python code applied to a dictionary of this language, they produce a frequency distribution for consonant-vowel sequences and uncover a near complementary distribution of *t* and *s*. Computational methods used with large data sets, and the probabilistic models they produce, have strong potential to help uncover previously unknown regularities and insights. For example, a simple combinatorics for English syllabic roles, ignoring frequency, predicts far more medial CCC clusters than actually occur. When viewed in terms of the joint probability of English onsets and codas, however, the actual frequencies of these clusters are surprisingly close to their expected frequencies (Pierrehumbert 1994).

These techniques allow researchers to make use of continuous variables to analyze phonological generalizations instead of traditional categorical variables. This has helped to uncover and explain processes that have been problematic in classic categorical approaches to phonology. For example, root co-occurrence restrictions have been difficult to predict with categorical restrictions on possible segments (Pierrehumbert 1993). Applied to the classic problem of consonant co-occurrence in Arabic roots, Frisch (1996) et seq. demonstrated the validity of continuous variables by showing how observed/expected values correlate with the phonological similarity of two sounds (see Frisch (2011) for more detailed discussion of similarity in phonology). Analyses using information theory have also shown the relevance of continuous variables, including intermediate types of phonological contrast that fall on a gradient scale of predictability (Hall 2009). Linguistically significant generalizations such as these, and the connections to theories of comprehension and

production that they support, are made possible by theories of phonology that recognize explicit roles for frequency and probability.

In a sense, gradient patterns and the frequency distributions supporting them have always attracted interest in phonology. Quantitative accounts of specific structures have supported serious inquiry of a host of topical problems. For example, Maddieson and Precoda (1992) investigate the role of CV frequencies in evaluating articulatory versus acoustic accounts of segment inventories. Contingency tables documenting consonant co-occurrence and vowel co-occurrence in roots have supported a host of theoretical pursuits in harmony and disharmony phenomena (Mester 1986; Yip 1989; Harlow 1991; Padgett 1995; Alderete and Finley 2016). Finally, a number of studies have focused on how stem phonotactics guides morphological operations (Zuraw 2000; Ernestus and Baayen 2003; Jones 2008).

Many phonologists are now pushing beyond these focused investigations of specific segmental frames and using the methods and measures discussed above to give more broad and open-ended accounts of probabilistic phonology. That is, statistical accounts of the distributions of phonological structure are becoming part of the primary linguistic description of many languages. For example, Leung et al. (2004) motivate a broad account of type and token frequencies in Cantonese, including descriptive statistics of segments, rimes, and tone, on the basis that these statistics give new insights into the language and are indispensable in psycholinguistic studies. Likewise, a number of researchers have documented gradient phonotactics for its own sake, giving descriptive accounts of segment distributions, segment combinations such as consonant co-occurrence, and prosodic patterns (Alderete and Bob 2005; Alderete and Bradshaw 2013; Rácz et al. 2016; Orzechowska and Ridouane 2018). The increased importance of the general quantitative accounts of the distributions of phonological structure is another indication of the importance of probabilistic phonology, in the empirical sense of this term.

4.2 Phonological variation

Phonological processes are variable in nature and are impacted by a host of linguistic and non-linguistic factors. The approaches to variable phonology discussed in section 3.1 address this problem head on by analyzing the probability of rule application and documenting the scope and magnitude of the conditioning factors in a standard re-write rule (Labov 1969; Cedergren and Sankoff 1974). Constraint-based approaches model variation with a stochastic component in constraint evaluation and a noise signal (Boersma and Hayes 2001; Coetzee and Pater 2011; Zuraw and Hayes 2017). Rather than positing variation on any structure in the context of a rule, these models limit variation to random permutations of constraints with some generality in phonological typology, after the role of these constraints in the grammar is established through learning. For illustrations and comparison across theories, see Coetzee and Pater (2011), who give an overview of the mechanics of Variable Rules and constraint-based analyses, and provide extensions to lexically governed variation. This work gives detailed analyses of *t*-deletion in English dialects in each of these models, though exemplar and information theoretic models are not examined.

As noted in section 3, information theoretic approaches explain the motivation for specific phonological processes like deletion and reduction by analyzing the intrinsic communicative value of form structure (Cohen Priva 2015; Cohen Priva 2017). In this approach, highly predictable forms, like the consonant [ŋ] after hearing the sequence [stændɪ_], contribute little to the signal, and may thus succumb to production constraints (e.g., reduction or deletion) with minimal signal loss. Exemplar and usage-based models of phonology also encode detailed information that may define the space of variable phonology, including phonological context and register. Though variation tends to focus on context-free phonetic attributes in these models, exemplars that integrate acoustically rich information at the word level can represent a wide range of variation (Drager 2011).

In terms of comparing these different approaches, as we discuss in section 5, we do not believe that these models are easy to distinguish on empirical grounds, either because the models are so similar that they make very similar empirical predictions, or simply because they focus on different kinds of variation. Instead, we think it is more fruitful to focus on the core assumptions, and pursue them vigorously as a way of validating specific theories. Some of the questions raised by core assumptions, for example, if grammar formed from universal constraints restricts variation internally the same way it does cross-linguistically, are fascinating to consider but far from concluded (section 5). However, one key assumption, namely whether phonological analysis must reference probabilities at all, has been pursued to conclusion. An earlier constraint-based approach based in Optimality Theory (Anttila 1997; Anttila 2007) differs from the models discussed

above in that it does not refer to probabilities or weights in the grammar. Instead, constraints in this model may be left unranked relative to other constraints, and the probability distribution of the variant forms is produced by permuting the rank for all unranked constraints and observing the percentage occurrence of the variants in all possible ranking outcomes. This approach has been successful with highly constrained systems, but it has also been criticized because it imposes empirically implausible constraints on the probabilities of variable outcomes (Boersma and Hayes 2001; Coetzee and Pater 2011), and it seems that the only way to address this problem is to introduce constraints that are not independently motivated. Thus, while there are some differences in how probability factors into analyses of phonological variation, it seems to be agreed that probability should be part of the grammatical account.

4.3 Experimental phonology

Many of the synchronic patterns described above have been investigated experimentally with the goal of testing the impact of frequency and probability distribution on novel word-forms (Solé et al. 2007; Kawahara 2011). A standard experimental procedure is to probe native speaker judgments on non-word stimuli (Coleman and Pierrehumbert 1997). These judgments reflect phonotactic probabilities (Frisch et al. 2000; Treiman et al. 2000; Bailey and Hahn 2001; Frisch and Zawaydeh 2001; Frisch and Stearns 2006; Albright 2009; Frisch and Brea-Spahn 2010), suggesting that speaker knowledge is gradient in nature and correlates with probability distributions found in the lexicon. While lexically-based approaches directly encode this knowledge as an emergent property of the lexicon (Plaut and Kello 1999; Frisch et al. 2000), harmony scores and related maximum entropy values used in constraint-based approaches are also correlated with gradient acceptability (Keller 2006; Hayes and Wilson 2008).

Experimental investigations have also tested the role of probability in phonological processes by manipulating the factors known to affect phonological alternations (Crosswhite et al. 2003; Ernestus and Baayen 2003; Ernestus and Baayen 2004; Hayes and Londe 2006; Zuraw 2007; Hayes et al. 2009; Moore-Cantwell 2016). For example, Ernestus and Baayen (2003) probed the neutralization of obstruent [voice] in Dutch by presenting subjects with forms that had an unvoiced final obstruent, and then asked them to produce a past tense form which revealed their guess as to the underlying [voice] specification. Responses correlated with the frequency distribution of [voice] neutralization in the lexicon, including the frequencies of the affected segment and the contexts for that segment. This work supports both a stochastic component in constraint-based grammar (Hayes and Londe 2006) and connectionist or analogical models that are sensitive to the phonological similarity structure in the lexicon (Ernestus and Baayen 2003; Alderete et al. 2013).

One of the goals of experimental phonology is to help provide psycholinguistic evidence in favor of a given theoretical model. For example, Hayes and Londe's (2006) wug test data is used to support Zuraw's (2000) model of the lexicon and a stochastic, constraint-based approach to probabilistic grammar. Much of the work that supports exemplar and usage-based models of phonology is also based on experimental evidence. However, many of these experiments are used to inform the debate on whether speakers store abstract representations of lexical items (Hay et al. 2006; McQueen et al. 2006; Tilsen 2009), rather than to inform which theoretical model best captures probabilistic data. Model comparison is typically based on two metrics: model fit and simplicity of assumptions. Generally, the various modeling approaches under comparison fit and capture experimental data relatively well (Ernestus and Baayen 2003; Coetzee and Pater 2011), suggesting that simplicity is often used as a deciding factor for arguing for a particular approach. For example, Ernestus and Baayen (2003) compare five different models, including Stochastic OT, VARBRUL, and an analogical model on the Dutch voicing neutralization data discussed above. While all five models capture the variation in the data well, Ernestus and Baayen (2003) argue for analogical models that require fewer parameters than constraint-based (OT) models.

As noted above, when comparing models of experimental phonology, they tend to have a good fit with the experimental data (e.g., Hayes and Londe (2006) and Ernestus and Baayen (2003)). However, judgments gleaned from experimental data do not always directly mirror the probabilities found in corpora, nor is it clear whether the probabilities should be fit to model within or between speaker variability, because modeling the average of a population of speakers may not reflect any individual's phonological knowledge. For example, Crosswhite et al. (2003) found that while Russian native speakers' stress assignments in nonce forms were in the same direction as those found in corpora, subjects' responses did not directly match lexical frequencies.

Other studies have documented different levels of acceptability for forms that have the same frequency (Moreton 2002), and that only a subset of the viable statistical regularities in the lexicon impacts native

speaker judgments (Becker et al. 2011; Moore-Cantwell 2013). A large scale study of syntactic acceptability in English also found that acceptability is gradient, but this acceptability does not reduce to probability (Lau et al. 2017). Thus, while many studies have shown a strong fit between experimental and corpus data, it is clear that probability alone cannot account for many facets of linguistic knowledge. More research is needed to better understand how probability distributions are used in speakers' knowledge. For example, probability distributions over some representations (e.g., voiced consonants) may yield different results than probability distributions over others (e.g., stressed vowels), and these distributions may interact in interesting ways (Kapatsinski 2012).

4.4 Phonological categorization

There is a significant amount of evidence that the production and perception of phonological categories are probabilistic (Pierrehumbert 2003b). First, even within a single speaker, for the same token, the phonetic properties of a given segment will vary from utterance to utterance, and this variation increases with individual speakers across a variety of utterances. One way to formalize the connection between production and perception in shaping phonological categories is through Bidirectional Optimality Theory (Boersma and Hamann 2008). In Bidirectional OT, perceptual changes result in different lexical representations, which can then be integrated to production systems. This back and forth gives rise to sound inventories that optimize constraints on both perception and production. Other constraints on vowel inventory can be drawn from dispersion theory (Lindblom 1986). Dispersion theory predicts the shape of vowel inventories in terms of dispersion across the acoustic space, and that these inventories should take up as much acoustic space as possible. This predicts, for example, that a three-vowel inventory will not include three high/front vowels, but will more likely include an inventory similar to [i, a, u] that takes up the majority of the acoustic space. Dispersion theory can be used to explain perceptual constraints on phonological behavior that is heavily influenced by vowel inventories such as vowel reduction (Padgett 2004; Padgett and Tabain 2005) and vowel harmony (McCollum 2018).

According to many exemplar models of phonology (e.g., Pierrehumbert (2001b); Pierrehumbert (2003a)), representations of phonetic categories form a "cloud" of the phonetic properties of individual utterances. This poses a challenge to the learner because there is a large amount of overlap within the phonetic realization of different sounds (e.g., the phonetic realization of [e] often overlaps with the phonetic realization of [ɛ]). However, the amount of overlap will vary depending on context and allophonic versus phonemic status. Maye et al. (2002) demonstrated that infants are more likely to infer two categories when the distribution resembles a bimodal distribution, and a single category when the distribution is unimodal. Bayesian analyses of categorization based on exemplar representations that track context, frequency, extent of overlap, and information from "higher levels" of abstraction (such as a word or morpheme) can allow the learner to infer, with measured success, the phonetic distributions of the categories in one's language (Shi et al. 2010). Information at higher levels of abstraction can also be used to provide information about contextual effects that constrain the phonetic realization of various categories, thus helping the learner to infer phonetic categories (Feldman et al. 2013).

4.5 Sound change

Probabilistic models of phonology that integrate frequency and language use (e.g., Bybee (2001), Zuraw (2003), Cohen Priva (2012), Hume and Mailhot (2013), Wedel et al. (2013), and Hay and Foulkes (2016)) can be used to address several important questions in sound change, such as why some sound changes are more common (or probable) within and across languages, and why some sound changes involve reduction, while others involve hyper-articulation or lengthening. Frequent items are more likely to be reduced, while confusable items are more likely to be hyper-articulated (Scarborough 2004). Because exemplar models encode phonetic details, frequency, and context, it is possible to model the diverse factors that contribute to a given sound change (Bybee 2000; Pierrehumbert 2001a). Wedel (2006), for example, models changing representations in terms of evolution, but is also compatible with analogies to lexical/neural activation, where the strengths of connections between items change as a result of the strength and frequency of use. In addition, functional pressures such as informativity can also be used to predict the probability and direction of sound change: more informative segments are more likely to be preserved or lengthened than less informative segments (Cohen Priva 2015).

4.6 Language development and segmentation

Given that phonological processes show probabilistic tendencies, it is not surprising that researchers have made great strides in understanding phonological development when looking at development through a probabilistic lens. Using a multidisciplinary approach to understanding the mechanisms that underlie speech segmentation, a tremendous amount of progress has been made in the last 25 years related to language development, particularly speech segmentation and phonotactic learning.

Speech segmentation is likely one of the first places where a human learner might make use of probability. As noted in Section 2, the probability of one syllable preceding another is higher within words than between words. After a very limited exposure to a speech stream, with no other cues for word segmentation other than transitional probability, infants were able to recognize the difference between words and partial words (Saffran et al. 1996; Aslin et al. 1998). In addition, learners may use knowledge of their phonological grammar to segment speech (Johnson and Jusczyk 2001), such as vowel harmony (Suomi et al. 1997; Vroomen et al. 1998) and phonotactic constraints (McQueen 1998). Newport and Aslin (2004) showed that learners were unable to segment speech in terms of non-adjacent dependencies based on syllables, but could segment speech based on non-adjacent consonants, suggesting that learners parse consonants and vowels differently, perhaps on separate "tiers". These results demonstrate the importance of distinguishing the level of representation that probabilities apply to. While transitional probabilities over segments can explain a large portion of speech segmentation, the progress that has been made would not be possible without integrating statistical information with more detailed linguistic cues, such as consonant/vowel status, prosody, coarticulation, and phonotactics (Daland and Pierrehumbert 2011). In fact, linguistic cues may be more reliable than statistical cues, as infants will segment speech in terms of linguistic cues over statistical cues when these cues are in conflict (Johnson and Jusczyk 2001). Models making use of Bayesian induction have been helpful for integrating phonotactic information into probabilistic induction of word boundaries (Goldwater and Johnson 2004; Daland and Pierrehumbert 2011).

Probabilistic approaches to learnability have also been used in phonotactic pattern learning, including Bayesian models (Shi et al. 2010; Feldman et al. 2013), and models that integrate statistical and linguistic cues (Chambers et al. 2003; White et al. 2008; Adriaans and Kager 2017; Schatz et al. 2021), demonstrating the role of probability in learning phonotactic constraints. Because statistical learning mechanisms are often based on transitional probabilities, such models of statistical learning of phonotactics may be compatible with finite-state based phonological grammars. In these grammars, phonological representations are based on transitions or precedence relationships (Heinz 2010). Learners may make use of transitional probabilities to help form abstract models of the grammar such as those used in finite-state models.

Probabilistic phonology is also consistent with theories of learning biases in phonology (Wilson 2006; Finley and Badecker 2007; Moreton 2008). Staubs (2014), for example, provides an error-driven learning account of the typology of stress systems using bigram statistics. These bigram statistics can be tied to both representational constraints on foot structure and probabilistic constraints on perception. In addition, language learners tend to produce more frequent patterns (e.g., CV syllables) earlier than less frequent patterns (e.g., CVC syllables) (Fikkert 1994; Fikkert 2007). Because frequency is often conflated with markedness, it is sometimes unclear how to interpret age of acquisition findings. It is also important to note that because certain developmental patterns (e.g., consonant harmony (Goad 2001)) do not reflect anything that the child may have heard, any theory of language development must allow for productions outside of the lexicon. A challenge to representing children's productions is that they may not reflect a child's lexical representations because children can distinguish their own mispronunciations from others' (e.g., a child will produce *fish* as *fis*, but will not accept others' mispronunciations (Smolensky 1996; Clark 2016)).

5. Discussion: Finding common ground

The breadth of the applications of probabilistic phonology reviewed above supports a consensus emerging in the field that frequency and probability are necessary components of phonological theory. Gradient phonotactic generalizations and variable phonological processes are attested in every language seriously investigated, and experimental testing of phonological alternations has also revealed a role for probability in a range of unrelated languages. Probabilistic patterns are not just a peripheral empirical dimension in phonology, or attested in a small number of cases. Rather, they constitute a core facet of phonological systems in general.

Probabilistic structure has also been developed across a range of theoretical models, which again underscores this emerging consensus. Probability has always been integral to psycholinguistic theories that refer to phonological structure (Jurafsky 2003), and several theories, such as exemplar and connectionist models, grew out of insights from psycholinguistics. Probability and frequency are also critical to the goals of information-theoretic phonology and the underlying measures they employ. While constraint-based theories do not necessarily require the use of lexical frequencies, constraint-based models can readily integrate probability and statistics with abstract constraints on representations.

One can view this integration of probabilistic structure on both empirical and theoretical levels as a kind of progress, because it reflects a shared view of the centrality of non-categorical structure in phonology. This point may not be appreciated by recent newcomers to the field, but the history of the development of phonological theory has not always been kind to this view (Hockett 1955; Halle 1962; Stanley 1967). Change has certainly occurred over the last 70 years (Pierrehumbert forthcoming).

Despite this progress, we contend that many problems still exist concerning precisely how assumptions about frequency and probability are implemented. In particular, given the diversity of perspectives we have reviewed above, the current state of the field can be characterized by a lack of cohesion that leads to a number of problems in the advancement of the science. The existence of many theories of probabilistic phonology leads to competing explanations of the same phenomena, yet most of these solutions have not been rigorously evaluated. Furthermore, attempts at comparative analysis suffer from problems that prevent an on-the-balance comparison. Finally, these models approach phonological analysis from very different base assumptions, with differing views about the existence of input-output mappings, roles for underlying representations, and treatment of allomorphy. These varying assumptions further inhibit progress because it becomes almost impossible to compare and build on work when the basic foundations vary so dramatically. This exacerbates the issue of silos in phonology, where researchers with shared interests but varied theoretical bents are unable to effectively communicate their findings to one another.

To address these concerns, we attempt to summarize some of the problems we see that lead to this lack of cohesion, and also suggest some approaches that we think aid in finding common ground.

5.1 Problems

We have presented several methods and applications for explaining probabilistic phonological patterns. One might be tempted to ask which one best captures the data in the most cognitively plausible way. Ideally, each different approach could be neatly compared, and the one with the best fit of the data, with the simplest assumptions would be deemed, the 'best'. However, comparing, for example, exemplar models with constraint-based generative models, is often problematic. First, different models are often used to explain different phenomena, leaving no common empirical ground for comparison. Second, when models can be compared, the focus is on descriptive adequacy, rather than explanatory power. Third, comparing "which model is best" may not be the appropriate question. Rather, a better question may be to try and understand the contribution from various different approaches to understand probabilistic effects.

It can be very difficult to compare approaches that have different goals because the problems of interest are often divergent. Researchers who use generative approaches are often interested in more abstract interactions within and across languages, like constraint interaction and typology, while researchers who use exemplar and information-theoretic approaches are typically interested in continuous phenomena, such as sub-lexical and sub-phonemic variation and talker-specific processes. When the goals of two competing approaches are so vastly different, an adequate comparison is extremely challenging. For example, if models based on information theory work to explain gradience in terms of historical change, but a maximum entropy model works to explain gradience as a learning problem, then it is not clear how to adequately compare each approach.

One criticism of generative models is that while they can account for probabilistic phenomena, it often is perceived as an afterthought, or add on, rather than a core part of the grammar. Because exemplar and information theoretic models (for example) are designed to account for probabilistic and gradient phenomena, they are better at accounting for intrinsically gradient phenomena. However, it is important not to conflate descriptive adequacy with explanatory adequacy. As Benus (2005) argues, an exemplar model of subphonemic variation in Hungarian vowel harmony is "descriptively adequate" but does not provide an explanation for the underlying principles that govern transparent vowels in vowel harmony more broadly. Model fit can also be misleading if predictive power and overfitting are not also assessed. Gorman (2013) examined the gradient

phonotactic well-formedness judgments from several experiments (e.g., Albright and Hayes (2003)), and found that the gradient models that were used to account for well-formedness judgments in those studies did no better than a simple baseline. This suggests that striving to find a "perfect fit" between human judgments and a statistical model may be addressing the wrong questions. Information theoretic accounts may also be problematic if they do not adequately operationalize communicative efficiency in a way that is both mathematically cogent and cognitively plausible. Model fit may also vary depending on what the data represent, and what the model is capturing. For example, a large corpus may be representative of a variation across a group of speakers, but may not necessarily represent variation in an individual speaker. A model that is meant to capture the cognitive processes and representations at the individual level may not benefit from a perfect fit from a corpus that does not represent any given individual.

One potential solution to this problem is to consider how insights from different approaches might inform the other, such as considering what representations might be obligatory, and at what level (e.g., which models require syllable structure or features, as shown in Daland et al. (2011)). Mathematically-based models, which are grounded in provable theorems, and often provide measures of precision and accuracy, can provide a template for generative theories. Mathematically-based models can, in turn, gain insights from generative models about the representations of phonological processes across language varieties. Understanding the different assumptions that are required to model a given phenomenon can lead to a better understanding of what is necessary and sufficient to understand gradience at different levels.

A different set of problems stem not from the actual object of study, but from the nature of the theories themselves. While some theories have such different assumptions that they are impossible to compare (e.g., exemplar theory and OT), other theories have such similar assumptions that the predictions are almost identical (e.g., MaxEnt Harmonic Grammar and Noisy Harmonic Grammar). For example, variations of constraint-based models make use of relatively similar mathematical foundations, and often have very similar predictions. Thus, MaxEnt Harmonic Grammar and Noisy Harmonic Grammar can both learn a probabilistic version of constraint weights. One difference is that MaxEnt HG defines a probability distribution over a candidate set, while Noisy HG produces a single optimal output. This difference leads to relatively minor predictions about the probabilities of harmonically bounded candidates and also different assumptions about the types of data models are exposed to (Hayes and Wilson 2008; Coetzee and Pater 2011). However, both methods for grammar learning produce very similar results, raising questions about what can be learned from different versions of Harmonic Grammar when the results are relatively similar.

Often times, the differences between approaches are related to entrenched theoretical philosophies, making it difficult to isolate predictions related to probabilistic phonology. For example, many theories of phonology like OT and HG take a Universal Grammar (UG) as the null hypothesis: all constraints are innate and universal, making the learning problem one of determining the appropriate ranking of constraints for a given language. However, it is not clear what bearing the role of UG in OT and HG have on probabilistic phonology. While such questions related to the theoretical architecture (Paolillo 2002; Coetzee and Pater 2011) are interesting, they are generally orthogonal to how to integrate probability in phonology. While theories like Harmonic Grammar have a commitment to UG, they can be loosened to admit language-particular constraints. For example, Hayes and Wilson (2008) allow for induction of constraints from a universal set of representations, raising questions about where UG might reside (e.g., in the constraints or the representations that build the constraints). Likewise, models that do not make use of UG, such as Variable Rules phonology, can be modified to limit the factors influencing free variation to cross-linguistically attested patterns, as is common in practice (Labov 2004).

While it can be a challenge to compare differences between versions of Harmonic Grammar because they are so similar, the problem of comparison can also occur when the assumptions are fundamentally different. Root co-occurrence restrictions in Arabic have been studied from a variety of perspectives, and would appear to be an ideal candidate to contrast different approaches. However, a comparison of two recent accounts of these restrictions, namely Coetzee and Pater (2008) and Frisch et al. (2004), reveals the difficulty in comparing even radically different theories. Coetzee and Pater (2008) trained a Harmonic Grammar (with 17 feature-specific and context-specific Place cooccurrence constraints, and one faithfulness constraint) on representative Arabic data using standard error-corrective learning methods. The result is a set of constraint weights that approximate the relative well-formedness of Arabic roots using the gradient measure of relative harmony, a measure derived from harmony by comparing the winning candidate with the next-best competitor.

Frisch et al. (2004), on the other hand, proposed that Arabic roots are subject to similarity avoidance, a gradient constraint against consonant pairs. This constraint predicts a gradient in which pairs with greater similarity (i.e., pairs of segments that share more phonological classes) are less well-formed. The Similarity Avoidance account is based on parallels in speech production, which exhibit a related avoidance of identical and similar segments.

When both approaches are compared in Coetzee and Pater (2008) using r^2 (a measure of model fit), the Harmonic Grammar account fares better than the Similarity Avoidance approach. For example, the fit of relative harmony and O/E (observed/expected) on the Harmonic Grammar account is 0.40 for all non-identical consonant pairs, but only 0.20 when fit with similarity on the Similarity Avoidance account. As pointed out in Frisch (2011), however, this comparison does not take into consideration the differences in the degrees of freedom afforded in each analysis. The linear regression analysis (that produces the r^2 variable) for both approaches introduces two degrees of freedom, because the attested data is fit by correlation with a predictor value. However, the Harmonic Grammar analysis has many additional "hidden" degrees of freedom because relative harmony derives from the weighted sum of 18 distinct constraints, and all of these constraints receive weight coefficients by a procedure for learning weights from a large corpus of Arabic roots. Clearly, these independent parameters, tied to specific combinations of place and manner classes, lead to a better fit because the disharmonic patterns are sensitive to them. The Similarity Avoidance account, on the other hand, is not trained on large data sets, and does not have a large number of independent parameters. Indeed, similarity-based accounts have avoided using comparable mechanisms, like the weighting of specific features in calculating similarity (Pierrehumbert 1993; Frisch et al. 2004), precisely because they introduce additional degrees of freedom and are therefore not a substantive test of their theory. For this reason, it is not clear that a universal set of weighted constraints (as proposed by the HG analysis) is necessarily a better approach to the Arabic data. While the Similarity Avoidance approach may not, on further analysis, have the same descriptive adequacy of the HG approach, it may have more explanatory power, as this approach is grounded in psycholinguistic principles of phonological processing.

5.2 Some productive trends

Given these issues, how do we find common ground? Some recent work has discussed modest modification to existing models that represent improvements. For example, Coetzee and Kawahara (2013) discuss introducing sociolinguistic variables to harmonic grammars of variable phonology as a way of addressing some of the social factors that are important to the original accounts of Variable Rules phonology. Another model improvement uses Bayesian inference to determine how evidence drawn from in-the-moment phonetic reduction and enhancement can lead to long-term phonological changes (Hall et al. 2018). By giving problems like sound mergers, which have been analyzed in exemplar phonology, a Bayesian analysis, this approach provides a framework for making explicit parallels between research in speech perception and phonological analysis (see also Flemming (2010)). While these certainly are positive developments, we believe the depth and extent of the problems discussed above requires more action to find common ground. We highlight below some productive trends that we think will advance the cause.

One trend involves focusing squarely on the underlying mechanism that drives a pattern, and then establishing all of the steps that link the mechanism to phonological analysis. In a sense, contemporary phonology is keenly focused on underlying mechanisms (e.g., Boersma (1998)), and constraint-based phonology in particular strives to provide phonetic and function motivation for phonological patterns. For example, in addition to its typological motivation, Pater (1999) gives explicit motivation for *NÇ, a constraint banning nasal+voiceless obstruent clusters, on articulatory grounds (the banned sequences require an unnaturally quick velic closure). While such motivations are important, it has yet to be established how the underlying motivation leads ultimately to attested synchronic patterns. In this case, how do we get from the articulatory pressures in post-nasal voicing to the typological patterns banning voicelessness after nasals? We think that establishing this link can help address the problems arising from the diversity of perspectives in probabilistic phonology.

This "following all the steps" approach can be illustrated with an example from dissimilation. Like Pater's (1999) articulatory account, the mechanism for dissimilation starts with a parallel between speech production and formal phonology. Dissimilation systems exhibit an avoidance of segment repetition and sequences of similar sounds, and the functional grounding of these sequences has been supported by research establishing similarity avoidance in language production (Shattuck-Hufnagel 1979; Cutler 1980). Thus, the

constraints that govern the Obligatory Contour Principle (OCP) and are commonplace in phonological analyses of dissimilation and root cooccurrence restrictions (Goldsmith 1976; McCarthy 1986) have a functional motivation in speech production research. This parallel between the properties of on-line language production and phonological systems is established in some detail in Frisch (1996 et seq). Frisch (1996) compared the phonological similarity of English consonant pairs with spreading activation (the proxy for similarity in Dell's (1986) model of language production), and found that indeed consonant similarity correlated well with activation values. This correlation supported research that shows how similarity avoidance in on-line production patterns can influence long-term lexical representations through talker-listener interactions (Frisch 2004; Martin 2007). In particular, Martin (2007) sketches an agent-based modeling account of how the mechanisms that underlie similarity avoidance can exert biases in learning, and result in the persistence of skewed dissimilatory patterns. While there are alternative functional accounts that focus on perceptual difficulty with segment repetition (Boersma 1998; Frisch 2004), Frisch's account is particularly promising in this regard, and it serves as a model for integrating psycholinguistic grounding with formal phonological analyses.

This focus on the underlying mechanism and establishing a connection to phonological analysis has many advantages. First, by linking phonology to an underlying mechanism, it can be argued that such an analysis provides a genuine explanation for the existence of a pattern. It answers *why* questions rather than just *how* questions. We also think that focusing on explicit mechanisms leads to more explicit and testable predictions (e.g., the role of distance in the production account of dissimilation (Frisch 2004)). Moreover, by establishing an explicit link between underlying behavior and phonological patterns, the different steps in the analysis provide solutions at different levels of analysis. Low-level production data can be investigated for the variable patterns that lead to dissimilation, and the outcomes of agent-based modeling lead to analyses of complete phonological systems. Finally, and perhaps most importantly, a focus on an underlying mechanism in a sense makes deciding on a particular theoretical model somewhat less important. In the case of dissimilation, once the research has established that similarity avoidance emerges from constraints on speech production, then formalization can be developed in a variety of ways, as long as they account for the central facts. For example, similarity avoidance can be accounted for in a variety of production models (Dell 1986; Vousden et al. 2000; Smolensky et al. 2014), and the phonological analysis can be formalized in a connectionist network (Alderete et al. 2013) or a harmonic grammar (Coetzee and Pater 2008). The choice is not terribly important if both demonstrably account for the emergence of dissimilation. In other words, once the underlying mechanism has been established, the formalization question is secondary to implementing a model that approximates the behavior consistent with that mechanism.

While model comparison can be a challenge when the assumptions are either so radically different that they cannot be compared, or so similar that the differences are trivial, there are cases where model comparison can yield interesting and insightful findings. For example, Daland et al. (2011) created multiple versions of different models (e.g., versions that included syllable structure, and versions without) in order to compare within and across model categories. Their results quite clearly showed a need for syllable structure. Because computational models require a host of choices about parameters, it can be helpful to demonstrate how different choices affect model fit across a variety of models. These choices can be formal characteristics, as well as more representational choices, such as syllabification, featural information, or prosody (Hayes and Wilson 2008; Daland and Pierrehumbert 2011).

Another productive trend involves considering probabilistic patterns from various levels of analysis and relationships among those levels. Phonological processes are often most clearly expressed through morphophonological alternations. As Pierrehumbert (2016) argues, morphophonology may be well-suited to serve as a model for a hybrid approach, because the purpose of morphophonology is to understand the relationship between words within and across paradigms, but that indexical information (e.g., class, gender) has varying effects on generalization in different phonological contexts. In this approach, it is not about comparing models to find "which is best", but integrating differing factors into a single model. To our knowledge, no one has yet proposed a hybrid model of morphophonology that incorporates elements from an exemplar model with elements from a generative model. While it might be possible to tease apart categorical effects from gradient effects in morphophonology, it is not yet clear what such a model might look like.

The Integrated Connectionist/Symbolic (ICS) cognitive architecture (Smolensky 2006; Smolensky and Legendre 2006) and further developments from that approach (Smolensky et al. 2014) might serve as a

blueprint for such a model. The ICS framework integrates Harmonic Grammar (Legendre et al. 1990) across levels, and the larger model is meant to serve as a hybrid between connectionist and symbolic cognitive architectures. The microlevel operates like a neural network with distributed representations and spreading activations, while the macrolevel makes use of symbolic computations. The bridge between these two levels is harmony, which serves as the well-formedness measure used in Harmonic Grammar and OT. An architecture that has room for gradient and abstract levels of representation, with connections in between, may help explain why phonology requires categorical, abstract representations, in addition to the many gradient properties discussed in this paper. The ICS framework may also bridge gaps between psycholinguistic models and generative models. For example, artificial neural network models of speech perception (McClelland and Elman 1986) and speech production (Dell 1986; Dell et al. 1993) allow for connections between different layers of the network (e.g., features at a lower level, phonemes at an intermediate level, and words at a higher level).

An integrated account of the gradient and the abstract would likely involve a "ladder of abstraction", where different effects can be accounted for at various levels of abstraction. For example, gradient representations of speech can be influenced by abstract indexation of social class (Munson et al. 2012). A ladder of abstraction might also make sense when discussing gradient effects in phonological processing. For example, continuous acoustic information at one level of abstraction could influence more abstract levels of representation, like syllable structure. Pierrehumbert (2016) argues that many of the gradient and talker-specific effects appear at the individual word level, while generalizations beyond known words (e.g., novel morphophonological alternations, wug tests, etc.) appear to trigger more abstract representations (Finley 2013). This raises important questions about the relationship between the lexicon and the grammar (see, e.g., Myers (2007)), in that the lexicon might store many gradient, talker-specific representations, but an abstract grammar can be applied to novel forms. If these abstract representations apply probabilistically, then it might explain why some gradient effects in phonology cannot be reduced to a list of exemplars or formant values.

6. Conclusion

In the last 25 years, there has been a surge of research exploring the role of probability in phonological representations. This research, based largely on corpus analyses and experimental data, has shown consistently that phonological knowledge cannot be represented solely in terms of binary grammaticality judgments. Rather, speaker knowledge of phonotactics, morphophonology, and prosodic phonology must have more flexible representations. What these representations look like, and how they might be learned, are still subject to debate. We have surveyed several approaches, including lexically-based models such as exemplar theory, computational approaches that make use of statistical modeling and information theory, and constraint-based models such as MaxEnt grammar. These approaches can be applied to a host of phenomena, including language variation and change, psycholinguistic data, and language learning and development. Because probability has become an increasingly integral part of "mainstream" phonology, it is important to find common ground among the many approaches to and applications of probabilistic phonology. This paper highlights some of the issues related to finding such common ground, including ideological commitment to particular approaches and applications, as well as differences in the fundamental questions related to phonological knowledge. While probability plays an important role in phonology, there is a risk for research to lose sight of the big questions in phonology, relating to representation and grammatical knowledge, in favor of a model with the best fit of a given data set. Future research will benefit from a focus on finding ways to integrate abstract representations with non-binary representations. We suggest that research is likely to find support for an integrated approach to probability in phonology, where probability applies differently at various levels of abstraction.

Acknowledgements

We are grateful to Queenie Chan, Stefan Frisch, Kathleen Hall, Paul Tupper, Jie Zhang and two anonymous Language and Linguistics reviewers for questions and comments on earlier drafts of this article. This research was supported in part by an Insight grant from the Social Science and Humanities Research Council of Canada (435-2014-0452).

References

Adriaans, F. W and René Kager. (2017). Learning novel phonotactics from exposure to continuous speech. *Laboratory Phonology* 8. 1-14.

Albright, Adam. (2009). Feature-based generalization as a source of gradient acceptability. *Phonology* 26. 9-41.

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

Alderete, John and Tanya Bob. (2005). A corpus-based approach to Tahltan stress. In Sharon Hargus and Keren Rice (eds.), *Athabaskan prosody*, 369-391. Amsterdam: John Benjamins.

Alderete, John and Mark Bradshaw. (2013). Samoan root phonotactics: Digging deeper into the data. *Linguistic Discovery* 11. 1-21.

Alderete, John and Sara Finley. (2016). Gradient vowel harmony in Oceanic. *Language and Linguistics* 17. 769-796.

Alderete, John and Paul Tupper. (2018). Connectionist approaches to generative phonology. In Anna Bosch and S. J. Hannahs (eds.), *The Routledge handbook of phonological theory*, 360-390. New York: Routledge.

Alderete, John, Paul Tupper and Stefan A. Frisch. (2013). Phonological constraint induction in a connectionist network: Learning OCP-place constraints from data. *Language Sciences* 37. 52-69.

Anttila, Arto. (1997). Deriving variation from grammar. In Frans Hinskens, Roeland van Hout and W. Leo Wetzels (eds.), *Variation, change, and phonological theory*. Amsterdam: Benjamins.

—. (2007). Variation and optionality. In Paul de Lacy (eds.), *The Cambridge handbook of phonology*, 519-536. Cambridge: Cambridge University Press.

Aslin, Richard, Jenny Saffran and Elissa Newport. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9. 321-324.

Aylett, Matthew and Alice Turk. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47. 31-56.

Bailey, Todd M and Ulrike Hahn. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44. 568-591.

Becker, Michael and Maria Gouskova. (2016). Surface-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry* 47. 391-425.

Becker, Michael, Nihan Ketrez and Andrew Nevins. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87. 84-125.

Beckman, Mary and Jan Edwards. (2010). Generalizing over lexicons to predict consonant mastery. *Laboratory Phonology* 1. 319-343.

Benus, Stefan. (2005). Dynamics and transparency in vowel harmony. New York: New York University. (Doctoral dissertation.)

Bird, Steven and Edward Loper. (2004). NLTK: The natural language toolkit. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*.

Blevins, James P. (2006). Word-based morphology. *Journal of Linguistics* 42. 531-573.

Bod, Rens. (2003). Introduction to elementary probability theory and formal stochastic language theory. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic linguistics*, 11-37. Cambridge, MA: The MIT Press.

Boersma, Paul. (1998). *Functional Phonology*. The Hague: Holland Academic Graphics.

Boersma, Paul and Silke Hamann. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25. 217-270.

Boersma, Paul and Bruce Hayes. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32. 45-86.

Brent, Michael R. and Timothy A. Cartwright. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61. 93-125.

Bybee, Joan. (2000). The phonology of the lexicon: Evidence from lexical diffusion. In Michael Barlow and Suzanne Kemmer (eds.), *Usage-based models of language*, 65-85. Stanford: CSLI.

—. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan and James L. McClelland. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22. 381-410.
Cedergren, Henrietta. (1973). Interplay of social and linguistic factors in Panama. Ithaca: Cornell University. (Doctoral dissertation.)
Cedergren, Henrietta and David Sankoff. (1974). Variable rules: Performance as a statistical reflection of competence. *Language* 50. 333-355.
Chambers, K.E., K.H. Onishi and C. Fisher. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition* 87. B69-B77.
Chandlee, Jane and Jeffrey Heinz. (2016). Computational phonology. In Mark Aronoff (eds.), *Oxford Research Encylcopedia of Linguistics*. Oxford: Oxford University Press.
Chomsky, Noam. (1957). *Syntactic structures*. The Hague: Mouton.
—. (1961). Some methodological remarks on generative grammar. *Word* 17. 219-239.
—. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.
Chomsky, Noam and Morris Halle. (1968). *The sound pattern of English*. New York: Harper & Row.
Chomsky, Noam and George A. Miller. (1963). Introduction to the formal analysis of natural languages. In R. Duncan Luce, Robert R. Bush and Eugene Galanter (eds.), *Handbook of mathematical psychology, vol. 2*, 269-321. New York: John Wiley.
Clark, Eve V. (2016). *First language acquisition*. Cambridge: Cambridge University Press.
Coetzee, Andries and Shigeto Kawahara. (2013). Frequency biases in phonological variation. *Natural Language and Linguistic Theory* 30. 47-89.
Coetzee, Andries and Joe Pater. (2008). Weighted constraints and gradient restrictions on place co-ccurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 84. 289-337.
—. (2011). The place of variation in phonological theory (2nd edition). In John Goldsmith, Jason Riggle and Alan Yu (eds.), *The handbook of phonological theory*, 401-431. Malden, MA: Blackwell.
Cohen Priva, Uriel. (2012). Sign and signal: Deriving linguistic generalizations from information utility. Palo Alto: Stanford University. (Doctoral dissertation.)
—. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6. 243-278.
—. (2017). Informativity and the actuation of lenition. *Language* 93. 569-597.
Cole, Jennifer. (2009). Emergent feature structures: Harmony systems in exemplar models of phonology. *Language Sciences* 31. 144-160.
Coleman, John and Janet Pierrehumbert. (1997). Stochastic phonological grammars and acceptability. In (eds.), *Computational phonology. Third meeting of the ACL special interest group, Association of Computational Linguistics*, 49-56: Association for Computational Linguistics.
Crosswhite, Catherine, John Alderete, Tim Beasley and Vita Markman. (2003). Morphological effects on default stress placement in novel Russian words. In Gina Garding and Mimu Tsujimura (eds.), *Proceedings of the West Coast Conference on Formal Linguistics 22*, 151-164. Somerville, MA: Cascadilla Press.
Cutler, Anne. (1980). Errors of stress and intonation. In Victoria Fromkin (eds.), *Errors in linguistic performance: Slips of tongue, ear, pen, and hand*, 67-80. New York: Academic Press.
Daelemans, Walter, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch. (2002). *TiMBL: Tilburg memory-based learner reference guide, version 4.2.* Tilburg: Computational Linguistics Tilburg University.
Daland, Robert. (2013). Variation in the input: A case study of manner class frequencies. *Journal of Child Language* 40. 1091-1122.
—. (2014). What is computational phonology? *Loquens* 1. 2386-2637.
—. (2015). Long words in maximum entropy phonotactic grammars. *Phonology* 32. 353-383.
Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis and Ingrid Norrmann. (2011). Explaining sonority projection effects. *Phonology* 28. 197-234.
Daland, Robert and Janet B Pierrehumbert. (2011). Learning diphone-based segmentation. *Cognitive Science* 35. 119-155.
Daland, Robert and Kie Zuraw. (2018). Loci and locality of informational effects on phonetic implementation. *Linguistic Vanguard* 4. 1-10.

Dell, Gary S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93. 283-321.
Dell, Gary S., Cornell Juliano and Anita Govindjee. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science* 17. 149-195.
Dinnsen, Daniel A. and Jan Charles-Luce. (1984). Phonological neutralization, phonetic implementation and individual differences. *Journal of Phonetics* 12. 49-60.
Drager, Katie K. (2011). Sociophonetic variation and the lemma. *Journal of Phonetics* 39. 694-707.
Durand, Jacques, Ulrike Gut and Gjert Kristoffersen (eds). 2014. *The Oxford handbook of corpus phonology*. Oxford: Oxford University Press.
Eddington, David. (2000). Spanish stress assignment within the analogical modeling of language. *Language* 76. 92-109.
—. (2004). Issues in modeling language processing analogically. *Lingua* 114. 849-871.
Eisner, Jason. (2002). Parameter estimation for probabilistic finite-state transducers. *Proceedings of the 40th annual meeting of the Association of Computational Linguistics*. 1-8.
Ernestus, Mirjam. (2011). Gradience and categoricality in phonological theory. In Marc van Oostendorp, Colin J. Ewen and Keren Rice (eds.), *The Blackwell companion to phonology*, 2115-2136. Oxford: Blackwell.
Ernestus, Mirjam and Harald Baayen. (2011). Corpora and exemplars in phonology. In John A. Goldsmith, Jason Riggle and Alan C. Yu (eds.), *The handbook of phonological theory (2nd edition)*, 374–400. Oxford: Wiley Blackwell.
Ernestus, Mirjam and R. Harald Baayen. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79. 5-38.
—. (2004). Analogical effects in regular past tense production in Dutch. *Linguistics* 42. 873-903.
Feldman, Naomi H., Thomas L. Griffiths, Sharon Goldwater and James L. Morgan. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review* 120. 751-778.
Feldman, Naomi H., Thomas L. Griffiths and James L. Morgan. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the Annual Meeting of the Cognitive Science Society, Vol 31*.
Fikkert, Paula. (1994). On the acquisition of prosodic structure. Leiden: Leiden University. (Doctoral dissertation.)
—. (2007). Acquiring phonology. Handbook of phonological theory. In Paul de Lacy (eds.), *Handbook of phonological theory*, 537-554. Cambridge: Cambridge University Press.
Finley, Sara. (2013). Generalization to unfamiliar talkers in artificial language learning. *Psychonomic Bulletin & Review* 20. 780-789.
Finley, Sara and William Badecker. (2007). Towards a substantively biased theory of learning. *Proceedings of the 33rd meeting of the Berkeley Linguistics Society*.
Flemming, Edward. (2010). Modeling listeners. In C. Fourgeron, B. Kühnert, M. D'Imperio and N. Vallée (eds.), *Laboratory phonology 10*, 587-606. Berlin: De Gruyter Mouton.
Frank, Michael C., Noah D. Goodman and Joshua B. Tenenbaum. (2007). A Bayesian framework for cross-situational word-learning. *Advances in Neural Information Processing Systems* 20.
Frisch, Stefan A. (1996). Similarity and frequency in phonology. Evanston: Northwestern University. (Doctoral dissertation.)
—. (2004). Language processing and segmental OCP effects. In Bruce Hayes, Robert Kirchner and Donca Steriade (eds.), *Phonetically-based phonology*, 346-371. Cambridge: Cambridge University Press.
—. (2011). Frequency effects. In Marc van Oostendorp, Colin J Ewen, Elizabeth Hume and Keren Rice (eds.), *The Blackwell companion to phonology*, 2137-2163. Malden, MA: Blackwell.
—. (2012). Phonotactic patterns in lexical corpora. In Abigail C. Cohn, Cécile Fougeron, Marie K. Huffman and Margaret E. Renwick (eds.), *The Oxford handbook of laboratory phonology*, 458-470. Oxford: Oxford University Press.
Frisch, Stefan A. and Maria Brea-Spahn. (2010). Metalinguistic judgments of phonotactics by monolinguals and bilinguals. *Laboratory Phonology* 1. 345-360.
Frisch, Stefan A., Nathan R. Large and David S. Pisoni. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42. 481-496.

Frisch, Stefan A., Janet Pierrehumbert and Michael B. Broe. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22. 179-228.
Frisch, Stefan A. and Adrienne M. Stearns. (2006). Linguistic and metalinguistic tasks in phonology: Methods and findings. In Gisbert Fanslow, Caroline Féry, Matthias Schlesewsky and Ralf Vogel (eds.), *Gradience in grammar: Generative perspectives*, 70-84. Oxford: Oxford University Press.
Frisch, Stefan A. and Bushra Zawaydeh. (2001). The psychological reality of OCP-Place in Arabic. *Language* 77. 91-106.
Garrett, Andrew and Keith Johnson. (2013). Phonetic bias in sound change. In Alan Yu (eds.), *Origins of sound change: Approaches to phonologization*, 51-97. Oxford: Oxford University Press.
Goad, Heather. (2001). Assimilation phenomena and initial constraint ranking in early grammars. In H.-J. A. Do, L. Dominguez and A. Johansen (eds.), *Proceedings of the 25th annual Boston University Conference on Language Development*, 307-318. Somerville, MA: Cascadilla Press.
Goldrick, Matthew. (2007). Connectionist principles in theories of speech production. In Gareth Gaskell (eds.), *The Oxford handbook of psycholinguistics*, 515-530. Oxford: Oxford University Press.
Goldrick, Matthew and Sheila Blumstein. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes* 21. 649-683.
Goldrick, Matthew and Robert Daland. (2009). Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. *Phonology* 26. 147-185.
Goldsmith, John. (1976). Autosegmental phonology. Cambridge, MA: MIT. (Doctoral dissertation.)
— (ed.) 1993. *The last phonological rule: Reflections on constraints and derivations*. Chicago: University of Chicago Press.
Goldsmith, John and Gary Larson. (1990). Local modeling and syllabification. In K. Deaton, M. Noske and M. Ziolkowski (eds.), *Proceedings of the 26th annual meeting of the Chicago Linguistics Society, Part 2.* Chicago: Chicago Linguistics Society.
Goldsmith, John and Jason Riggle. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language and Linguistic Theory* 30. 859-896.
Goldwater, Sharon, Thomas L. Griffiths and Mark Johnson. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112. 21-54.
Goldwater, Sharon and Mark Johnson. (2003). Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson and Östen Dahl (eds.), *Proceedings of the Stockholm workshop on variation within Optimality Theory*. Stockholm: Stockholm University.
—. (2004). Priors in Bayesian learning of phonological rules. *Proceedings of the 7th meeting of the Association of Computational Linguistics: Current themes in computational phonology and morphology*. 35-42.
Gorman, Kyle. (2013). Generative phonotactics. Philadelphia: University of Pennsylvania. (Doctoral dissertation.)
Guy, Gregory R. (1992). Contextual conditioning in variable lexical phonology. *Language Variation and Change* 3. 223-239.
Guy, Gregory R. (1991). Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3. 1-22.
Hall, Kathleen. (2009). A probabilistic model of phonological relationships from contrast to allophony. Columbus, OH: Ohio State University. (Doctoral dissertation.)
—. (2012). Phonological relationships: A probabilistic model. *McGill Working Papers in Linguistics* 22. 1-14.
Hall, Kathleen, Blake Allen, Michael Fry, Scott Mackie and Michael McAuliffe. (2016). Phonological CorpusTools. *14th Conference for Laboratory Phonology*.
Hall, Kathleen, Elizabeth Hume, T. Florian Jaeger and Andrew B. Wedel. (2018). The role of predicability in shaping phonological patterns. *Linguistic Vanguard* 4. 5200170027.
Halle, Morris. (1962). Phonology in generative grammar. *Word* 18. 54-72.
Hare, Mary. (1990). The role of similarity in Hungarian vowel harmony: A connectionist account. In Noel Sharkey (eds.), *Connectionist natural language processing*, 295-322. Oxford: Intellect.

Harlow, Ray. (1991). Consonant dissimilation in Maori. In Robert Blust (eds.), *Currents in Pacific Linguistics: Papers on Austronesian languages and ethnolinguistics in honour of George W. Grace. Pacific Linguistics Series C, no. 117*, 117-128. Canberra: Australian National University.
Hashimoto, Daiki. (2021). Probabilistic reduction and mental accumulation in Japanese: Frequency, contextual predictability, and average predictability. *Journal of Phonetics* 87. 101061.
Hay, Jennifer and Paul Foulkes. (2016). The evolution of medial /t/ over real and remembered time. *Language* 92. 298-330.
Hay, Jennifer, Aaron Nolan and Katie K. Drager. (2006). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review* 23. 351-379.
Hayes, Bruce and Zsuzsa Cziráky Londe. (2006). Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23. 59-104.
Hayes, Bruce, Péter Siptár, Kie Zuraw and Zsuzsa Londe. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85. 822-863.
Hayes, Bruce and Colin Wilson. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39. 379-440.
Heinz, Jeffrey. (2010). Learning long-distance phonotactics. *Linguistic Inquiry* 41. 623-661.
Hockett, Charles Francis. (1955). *A manual of phonology*. Baltimore: Waverly Press.
Hume, Elizabeth. (2008). Markedness and the language user. *Phonological Studies* 11. 83-98.
Hume, Elizabeth and Frédéric Mailhot. (2013). The role of entropy and surprisal in phonologization and language change. In Alan Yu (eds.), *Origins of sound change: Approaches to phonologization*, 29–47. Oxford: Oxford University Press.
Jarosz, Gaja. (2006). Rich lexicons and restrictive grammars: Maximum likelihood learning in Optimality Theory. Baltimore: Johns Hopkins University. (Doctoral dissertation.)
Johnson, E. K. and P. W. Jusczyk. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language* 44. 548-556.
Johnson, Keith. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34. 485-499.
Jones, ʻŌiwi Parker. (2008). Phonotactic probability and the Māori passive: A computational approach. *Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2008)*. 39-48.
Jurafsky, Dan. (2003). Probablistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probablistic linguistics*, 39-95. Cambridge, MA: The MIT Press.
Kapatsinski, Vsevolod. (2012). What statistics do learners track? Rules, constraints or schemas in (artificial) grammar learning. In S Gries and D Divjak (eds.), *Frequency effects in language learning and processing*, 53-82. Berlin: Mouton de Gruyter.
Katz, Jerrold J. (1964). Semi-sentences. In Jerry A. Fodor and Jerrold J. Katz (eds.), *The structure of language: Readings in the philosophy of language*, 400–416. Englewood Cliffs, N.J.: Prentice-Hall.
Kawahara, Shigeto. (2011). Experimental approaches in theoretical phonology. In Marc van Oostendorp, Colin J Ewen, Elizabeth Hume and Keren Rice (eds.), *The Blackwell companion to phonology. Vol. 4, Phonological interfaces*. Malden, MA: Wiley-Blackwell.
Keller, Frank. (2000). Gradience in grammar: Experimental and computational aspects of degrees of grammaticality: University of Edinburgh. (Doctoral dissertation.)
—. (2006). Linear Optimality Theory as a model of gradience in grammar. In Fanselow Gisbert, Caroline Féry, Matthias Schlesewsky and Ralf Vogel (eds.), *Gradience in grammar: Generative perspectives*, 270–287. Oxford: Oxford University Press.
Kirov, Christo and Colin Wilson. (2013). Bayesian speech production: Evidence from latency and hyperarticulation. *Proceedings of the annual meeting of the Cognitive Science Society 35*. 788-793.
Labov, William. (1969). Contraction, deletion, and inherent variablity of the English copula. *Language* 45. 715-762.
—. (2004). Quantitative analysis of linguistic variation. 2nd edition. In Ulrich Ammon, Norbert Dittmer, Klaus J. Mattheier and Peter Trudgill (eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society. 2nd edition*, 6-21. Berlin: De Gruyter.

Laks, Bernard. (1995). A connectionist account of French syllabification. *Lingua* 95. 51-76.
Lau, Jey Han, Alexander Clark and Shalom Lappin. (2017). Grammaticality, acceptability, and probability: A probablistic view of linguistic knowledge. *Cognitive Science* 41. 1202-1241.
Legendre, Géraldine, Yoshiro Miyata and Paul Smolensky. (1990). Can connectionism contribute to syntax? Harmonic Grammar, with an application. In M. Ziolkowski, M. Noske and K. Deaton (eds.), *Proceedings of the 26th Regional Meeting of the Chicago Linguistic Society*, 237-252. Chicago: Chicago Linguistic Society.
Leung, Man Tak, Sam-Po Law and Suk-Yee Fung. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, and Computer* 36. 500-505.
Lindblom, Björn. (1986). Phonetic universals in vowel systems. In John J. Ohala and Jeri J. Jaeger (eds.), *Experimental phonology*, 13-44. Orlando: Academic Press.
Maddieson, Ian and Kristin Precoda. (1992). Syllable structure and phonetic models. *Phonology* 9. 45-60.
Martin, Andy. (2007). The evolving lexicon. Los Angeles: University of California, Los Angeles
Maye, Jessica, Janet F. Werker and LouAnn Gerken. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82. B101-B111.
McCarthy, John J. (1986). OCP Effects: Gemination and antigemination. *Linguistic Inquiry* 17. 207-263.
McClelland, James L. and Jeffrey Elman. (1986). The TRACE model of speech perception. *Cognitive Psychology* 18. 1-86.
McCollum, Adam. (2018). Vowel dispersion and Kazakh labial harmony. *Phonology* 35. 287-326.
McQueen, James M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language* 39. 21-46.
McQueen, James M., Anne Cutler and D. Norris. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science* 30. 1113-1126.
Mester, Ralf-Armin. (1986). Studies in tier structure: University of Massachusetts, Amherst. (Doctoral dissertation.)
Moore-Cantwell, Claire. (2013). Over- and under-generalization in derivational morphology. In Stefan Keine and Shayne Sloggett (eds.), *NELS 42: Proceedings of the 42nd meeting of the North East Linguistic Society*, 41-54. Amherst, MA: Graduate Linguistic Student Association.
—. (2016). The representation of probabilistic phonological patterns: Neurological, behavioral, and computational evidence from the English stress system. Amherst, MA: University of Massachusetts, Amherst. (Doctoral thesis.)
Moreton, Elliott. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition* 84. 55-71.
—. (2008). Analytic bias and phonological typology. *Phonology* 25. 83-127.
Munson, Benjamin, Jan Edwards and Mary Beckman. (2012). Phonological representations in language acquisition: Climbing the ladder of abstraction. In Abigail C. Cohn, Cécile Fougeron, Marie K. Huffman and Margaret E. L. Renwick (eds.), *The Oxford handbook of laboratory phonology*, 288-309. Oxford: Oxford University Press.
Myers, James. (2007). Linking data to grammar in phonology: Two case studies. *Concentric: Studies in Linguistics* 33. 1-22.
—. (2012). Testing phonological grammars with lexical data. In James Myers (eds.), *In search of grammar: Empirical methods in linguistics*, 141-176. Taipei: Language and Linguistics.
Myers, James and Jane Tsay. (2005). The processing of phonological acceptability judgments. Paper presented to the Proceedings of Symposium on 90-92 NSC Projects, Taipei, Taiwan, 2005.
Newport, Elissa and Richard Aslin. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology* 48. 127-162.
Norris, Dennis and James M. McQueen. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115. 357-395.
Nosofsky, Robert M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental psychology: General* 115. 39-57.
Orzechowska, Paula and Rachid Ridouane. (2018). The structure of vowelless verbal roots in Tashlhiyt Berber. Paper presented to the 9th International Conference on Speech Prosody 2018, Poznan, Poland, 2018.
Padgett, Jaye. (1995). *Stricture in feature geometry*. Stanford, CA: CSLI Publications.

—. (2004). Russian vowel reduction and Dispersion Theory. *Phonological Studies* 7. 81-96.
Padgett, Jaye and Marija Tabain. (2005). Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica* 62. 14-54.
Paolillo, John C. (2002). *Analyzing linguistic variation: Statistical models and methods*. Stanford, CA: CSLI Publications.
Pater, Joe. (1999). Austronesian nasal substitution and other NC effects. In René Kager, Harry van der Hulst and Wim Zonneveld (eds.), *The prosody morphology interface*, 310-343. Cambridge, MA: Cambridge University Press.
Pierce, John R. (1961). *An introduction to information theory: Symbols, signals, and noise*. New York: Dover.
Pierrehumbert, Janet. (1993). Dissimilarity in the Arabic verbal roots. In (eds.), *NELS 23*, 367-381.
—. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. In Patricia A. Keating (eds.), *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, 168-188. Cambridge: Cambridge University Press.
—. (2001a). Exemplar dynamics: Word frequency, lenition, and contrast. In Joan L Bybee and P Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137-158. Amsterdam: John Benjamins.
—. (2001b). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes* 16. 691-698.
—. (2003a). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 42. 115-154.
—. (2003b). Probabilistic phonology: Discrimation and robustness. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probability theory in linguistics*, 177-228. Cambridge, MA: The MIT Press.
—. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics* 2. 33-52.
—. (forthcoming). 70+ years of probablistic phonology. In (eds.), *The Oxford handbook of the history of phonology*. Oxford: Oxford University Press.
Plaut, David C. and Christopher T. Kello. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In Brian MacWhinney (eds.), *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates, Ltd.
Pluymaekers, Mark, Mirjam Ernestus and R. Harald Baayen. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62. 146-159.
Prince, Alan and Paul Smolensky. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Malden, MA: Blackwell.
Prince, Alan and Bruce Tesar. (2004). Learning phonotactic distributions. In René Kager and Joe Pater (eds.), *Fixing priorities: Constraints in phonological acquisition*, 245-291. Cambridge: Cambridge University Press.
Rácz, Péter, Jennifer Hay, Jeremy Needle, Jeanette King and Janet B Pierrehumbert. (2016). Gradient Māori phonotactics. *Te Reo* 59. 3-21.
Riggle, Jason. (2004). Generation, recognition, and learning in finite-state Optimality Theory. Los Angeles: University of California, Los Angeles. (Doctoral dissertation.)
Rose, Yvan, Brian MacWhinney, Rodrique Byrne, Gregory Hedlund, Keith Maddocks, Philip O'Brien and Todd Wareham. (2006). Introducting *Phon*: A software solution for the study of phonological acquisition. In (eds.), *Proceedings of the Boston University Conference on Language Development*, 489-500.
Saffran, Jenny, Richard Aslin and Elissa Newport. (1996). Statistical learning by 8-month-old infants. *Science* 274. 1926-1928.
Saffran, Jenny R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science* 12. 110–114.
Sankoff, David, Sali Tagliamonte and Eric Smith. (2005). GoldVarb X: A variable rule application for Macintosh and Windows.
Scarborough, Rebecca. (2004). Coarticulation and the structure of the lexicon: University of California, Los Angeles. (Doctoral dissertation.)

Schatz, Thomas, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao and Emmanuel Dupoux. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences of the United States of America* 118. e2001844118.
Shannon, Claude E. and Warren Weaver. (1949). *The mathematical theory of communication*. Urbana-Champaign: University of Illinois Press.
Shattuck-Hufnagel, Stefanie. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Copper and E. C. T. Walker (eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*, 295-342. Hillsdale, NJ: Erlbaum.
Shaw, Jason and Shigeto Kawahara. (2018). Predictability and phonology: Past, present, and future. *Linguistic Vanguard* 4. 20180042.
Shi, L., Thomas L. Griffiths, Naomi H. Feldman and A. N. Sanborn. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review* 17. 443-464.
Skousen, Royal. (1989). *Analogical modeling of language*. Dordrecht: Kluwer.
—. (1992). *Analogy and structure*. Dordrecht: Kluwer Academic.
—. (1995). Analogy: a non-rule alternative to neural networks. *Rivista di Linguistica* 7. 213-232.
Smolensky, Paul. (1988). On the proper treatment of connectionism. *Brain and Behavioral Sciences* 11. 1-23.
—. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry* 27. 720-731.
—. (2006). Harmony in linguistic cognition. *Cognitive Science* 30. 779-801.
Smolensky, Paul, Matthew Goldrick and Donald Mathis. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science* 38. 1107-1138.
Smolensky, Paul and Géraldine Legendre. (2006). *The harmonic mind. From neural computation to optimality theoretic grammar*. Cambridge, MA: The MIT Press.
Solé, Maria-Josep, Patrice Speeter Beddor and Manjari Ohala. (2007). *Experimental approaches to phonology*. Oxford: Oxford University Press.
Sorace, Antonella and Frank Keller. (2005). Gradience in linguistic data. *Lingua* 115. 1497-1524.
St. John, M. F. and James L. McClelland. (1988). Learning and applying contextual constraints in sentence comprehension. Technical Report AIP - 39 (ed.) University of Pittsburgh Carnegie Mellon University.
Stanley, Richard. (1967). Redundancy rules in phonology. *Language* 43. 393-436.
Staubs, Robert D. (2014). Computational modeling of learning biases in stress typology: University of Massachusetts, Amherst. (Doctoral dissertation.)
Suomi, Kari, James M. McQueen and Anne Cutler. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language* 36. 422-444.
Tesar, Bruce and Paul Smolensky. (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
Thomas, Michael S. C. and James L. McClelland. (2008). Connectionist models of cognition. In Ron Sun (eds.), *Cambridge handbook of computational psychology*, 23-58. Cambridge: Cambridge University Press.
Tilsen, Sam. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics* 37. 276-296.
Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff and Margo Bowman. (2000). English speakers' sensitivity to phonotactic patterns. In Michael B Broe and Janet B Pierrehumber (eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*, 269-282. Cambridge: Cambridge University Press.
Tupper, Paul. (2015). Exemplar dynamics and sound merger in language. *SIAM Journal on Applied Mathematics* 75. 1469-1492.
Turnbull, Rory. (2015). Assessing the listener-oriented account of predictability-based phonetic reduction. Columbus, OH: Ohio State University. (Doctoral dissertation.)
Vousden, Janet I., Gordon D.A. Brown and Trevor A. Harley. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology* 41. 101-175.
Vroomen, J., J. Tuomainen and B. de Gelder. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language* 38. 133-149.

Warner, Natasha, Allard Jongman, Joan A. Sereno and Rachèl Kemps. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics* 32. 251–276.
Wedel, Andrew B. (2003). Self-organization and categorical behavior in phonology. *Proceedings of the Berkeley Linguistics Society 29*. 611-622.
—. (2006). Exemplar models, evolution and language change. *The Linguistic Review* 23. 247-274.
Wedel, Andrew B., Abby Kaplan and Scott Jackson. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 128. 179-186.
White, Katherine S., Sharon Peperkamp, Cecilia Kirk and James L. Morgan. (2008). Rapid acquisition of phonological alternations by infants. *Cognition* 107. 238-265.
Wilson, Colin. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30. 945-982.
Wilson, Colin and Lisa Davidson. (2013). Bayesian analysis of non-native cluster production. *Proceedings of the Northeastern Linguistics Society, Vol. 40*. 265-278.
Xu, Fei and Josh Tenenbaum. (2007). Word learning as Bayesian inference. *Psychological Review* 114. 245-272.
Yip, Moira. (1989). Feature geometry and cooccurrence restrictions. *Phonology* 6. 349-374.
Zuraw, Kie. (2000). Patterned exceptions in phonology. Los Angeles: University of California, Los Angeles. (Doctoral dissertation.)
—. (2003). Probability in language change. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic linguistics*, 139-176. Cambridge, MA: The MIT Press.
—. (2007). The role of phonetic knowledge in phonotactic patterning: Corpus and survey evidence from Tagalog infixation. *Language* 83. 277-316.
Zuraw, Kie and Bruce Hayes. (2017). Intersecting constraint families: An argument for Harmonic Grammar. *Language* 93. 497-548.

Authors' address:
John Alderete (corresponding author):
Department of Linguistics
Simon Fraser University
8888 University Drive
Burnaby, V5A 1S6, Canada

alderete@sfu.ca

Publication history

Date received: 29 February 2020
Date accepted: 30 June 2021