ANNUAL REVIEWS

Annual Review of Linguistics

Computational Modeling of Phonological Learning

Gaja Jarosz

Department of Linguistics, University of Massachusetts, Amherst, Massachusetts 01003-1100, USA; email: jarosz@linguist.umass.edu

Annu. Rev. Linguist. 2019. 5:67–90

First published as a Review in Advance on August 20, 2018

The *Annual Review of Linguistics* is online at linguist.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-011718-011832

Keywords

learning, phonology, computational linguistics, statistical learning, hidden structure

Abstract

Recent advances in computational modeling have led to significant discoveries about the representation and acquisition of phonological knowledge and the limits on language learning and variation. These discoveries are the result of applying computational learning models to increasingly rich and complex natural language data while making increasingly realistic assumptions about the learning task. This article reviews the recent developments in computational modeling that have made connections between fully explicit theories of learning, naturally occurring corpus data, and the richness of psycholinguistic and typological data possible. These advances fall into two broad research areas: (*a*) the development of models capable of learning the quantitative, noisy, and inconsistent patterns that are characteristic of naturalistic data and (*b*) the development of models with the capacity to learn hidden phonological structure from unlabeled data. After reviewing these advances, the article summarizes some of the most significant consequent discoveries.

1. INTRODUCTION

Hidden structure: any abstract representation that underlies linguistic knowledge but is not directly observable in the learning data, such as metrical footing, underlying representations, and exceptionality diacritics

Recent advances in computational modeling of language learning have had a transformative impact on the field of phonology. These developments have made it possible to test and compare formally precise theories of learning and linguistic endowment while making increasingly realistic assumptions about the nature of the learning data and the learning task. Modeling has led to discoveries about how language knowledge is represented, how it is acquired, and the limits that exist on learning and variation. These discoveries would not have been possible without the formalization of the connection between natural language input and linguistic behavior that recent computational models of learning provide. This link has yielded new methods for testing theoretical assumptions by comparing the predictions of computational models with measurable linguistic behavior in psycholinguistic experiments, typological evidence, and empirical observations about language change and loanword adaptation.

One concrete shift in the field is methodological: Computational modeling has become an essential tool of modern phonological research. It complements the rise of experimental research on phonological knowledge and learning and the increase in available linguistic databases, both of which provide a rich and complex empirical base for developing and evaluating learning models and phonological theories. The mutually informing link between computational modeling and these growing empirical resources arose from modeling developments that can be broadly classified into two areas of research.

First, the development of learning models that can deal with the quantitative, variable, and inconsistent patterns that are characteristic of naturalistic data has made it possible to apply and test learning models on data representative of the linguistic experience of human language learners. While simulations with toy data that abstract from the irregularities of natural language are often an essential step in the development of new computational models, more realistic assumptions about the nature of the linguistic input permit greater confidence that resulting conclusions are applicable to human language learning. This is especially true when making claims about the sufficiency or insufficiency of the language input to support learning of some linguistic property or generalization—these questions can only be answered by examining the distribution and nature of the evidence in naturally occurring data. Likewise, only through detailed comparison of the quantitative patterns in natural language data and the generalizations that learners infer on the basis of those data can systematic biases can be fully understood. Section 2 reviews the most significant recent developments that have made it possible to model learning of phonology from naturalistic corpus data, arguing that these capabilities require the use of frequency-sensitive learning approaches such as those inherent to statistical learning models.

The second area of research in computational modeling involves the learning of hidden linguistic structure, representations that learners must infer but that cannot be directly observed in the learning data. Depending on theoretical assumptions, hidden structure in phonology may include metrical feet; underlying representations (URs); syllables; moraic structure; autosegmental associations; derivational ordering; word and other prosodic boundaries; and even the constraints, rules, and features themselves if they are not innately specified to the learner. Since children learn language without direct access to hidden representations, the capacity to learn these representations is essential to making realistic assumptions about the learning task. How these representations are inferred and how their learning interacts can only be understood via the development of explicit learning models capable of learning from incomplete and massively ambiguous data. Section 3 reviews significant discoveries in this area, arguing that statistical methods and other frequency-sensitive approaches have also been crucial to progress on hidden structure learning.

Finally, Section 4 reviews discoveries that have resulted from the application of frequency-sensitive models to psycholinguistic and typological questions. A recurring theme in many of

these studies is the fundamental question of nature versus nurture. What is the precise balance of experience sensitivity and innate predisposition that accounts for human learners' generalization from limited exposure to ambiguous, incomplete, and inconsistent natural language input? In what ways do learners systematically diverge from their language experience, and can these learning biases account for observed restrictions in language typology and language change? Evaluating models on their abilities to account for human learning and generalization is essential to answering these questions, providing a strict litmus test that has already revealed subtle complexities and strong constraints on the language acquisition device.

2. LEARNING QUANTITATIVE GENERALIZATIONS

Perhaps the most interesting and challenging aspect of modeling language acquisition is understanding how learners generalize from data that are inconsistent and incomplete. This section discusses the challenge posed by inconsistency, while Section 3 focuses on incompleteness, but these are two sides of the same coin: ambiguity. Ambiguity means that there are multiple interpretations or multiple analytic decisions that the learner could make to account for the same data. Understanding how learners disambiguate between the wealth of possible analyses of the same inconsistent, incomplete data gets at the very essence of language learning. To explain the choices learners make, it is necessary to make fully explicit how learners balance various considerations against one another and how they integrate various sources of information. Modeling acquisition from ambiguous data also provides the best opportunity to observe and formalize pressures that may bias learners' decisions toward phonetically natural, typologically common, and more systematic generalizations.

Inconsistencies in natural language data take many forms. The child learning their first language is not told which data tokens are errors that should be ignored, or which examples are exceptions to the general patterns they must infer. Language acquisition is robust enough to detect general patterns in the face of a few exceptions. Language acquisition must also be flexible enough to detect and differentiate these occasional divergences from the systematic variability that arises when the realizations of individual words or morphemes vary probabilistically and unpredictably in the same phonological environment. Speakers' knowledge of such free variation includes not only the categorical restrictions on the observed variability but also how the rate of variation depends on various phonological factors (for reviews, see Anttila 2007, Coetzee & Pater 2011). In the domain of gradient phonotactics, speakers show sensitivity to generalizations of varying degrees of productivity, and this sensitivity reflects quantitative properties of the language data, such as the probability of sound co-occurrences and the (under)attestation of certain sound combinations (Bailey & Hahn 2001, Coleman & Pierrehumbert 1997, Frisch et al. 2000, Hayes & Wilson 2008). Another type of inconsistency often found in natural language phonologies arises when lexical classes partition the lexicon into strata, each associated with distinct constellations of phonological processes and properties (Inkelas et al. 1997, Itô & Mester 1999). Finally, in patterned exceptionality, speakers have knowledge of language-wide quantitative trends while simultaneously encoding the fixed behavior of particular morphemes or morpheme combinations (Becker et al. 2011, Ernestus & Baayen 2003, Gouskova & Becker 2013, Hayes & Londe 2006, Zuraw 2000). In all of these cases, the learner is faced with patterns in which phonologically similar words or morphemes behave inconsistently in the same phonological environments.

The following subsections address these various forms of inconsistency, reviewing the approaches that have been developed to cope with them.[1] Addressing these inconsistencies requires

Ambiguity: when the learning data are, either locally or globally, compatible with a range of distinct analyses that the learner must navigate and choose between

Free variation: when a word or morpheme can be realized in multiple ways in the same environment; the choice of variants may be statistically conditioned by systematic phonological factors, but the variation is not entirely predictable

Gradient phonotactics: knowledge of legal and likely sound combinations that make up words in a language

Lexical classes: a partition of the lexicon into disjoint sets, each associated with a distinct constellation of phonological properties and/or processes

Patterned exceptionality: when systematic phonological factors statistically condition phonological variation in the aggregate across the lexicon but individual words exhibit fixed behavior

[1]Inconsistency can also arise through phonetic (Boersma 2011, Pierrehumbert 2001) and phonological (Legendre et al. 2006, Smolensky & Goldrick 2016) gradience.

sensitivity to quantitative properties of the data; therefore, much of this section focuses on ways in which computational models make use of quantitative information.

2.1. Preliminaries

Many of the models that have been developed to cope with quantitative phonological generalizations rely on probabilistic extensions of Optimality Theory (OT; Prince & Smolensky 2004) or Harmonic Grammar (HG; Legendre et al. 1990, Smolensky & Legendre 2006). Stochastic OT (Boersma 1997, Boersma & Hayes 2001), Noisy HG (Boersma & Pater 2016), and Maximum Entropy HG (MaxEnt; Goldwater & Johnson 2003, Jäger 2007, Johnson 2002, Wilson 2006) are three common probabilistic extensions of these frameworks (see also Jarosz 2015). Each of these frameworks encodes a stochastic grammar that assigns conditional probabilities to surface realizations of a given UR. This section illustrates these approaches using the MaxEnt model as an example (for in-depth comparisons, see Hayes 2017, Smith & Pater 2017).

Probabilistic constraint grammars formalize phonological mappings in terms of interactions of violable constraints, their language-specific prioritization, and the optimization that determines which among a set of candidate pronunciations is selected as the surface realization of a given UR. In MaxEnt (and HG), constraints are numerically weighted, and these weights are multiplied by the constraint violations incurred by each candidate and then summed to determine each candidate's overall harmony:

$$H(x,y) = \sum_{c \in C} w_c v_c(x,y).$$

The harmony $H(x,y)$ of an input–output pair (x,y) is the summation over all constraints $c \in C$ of the product of the weight of each constraint w_c and the number of violations $v_c(x,y)$ assigned to (x,y) by that constraint. Violations $v_c(x,y)$ are usually expressed as negative integers and weights w_c as nonnegative real values so that overall harmony is a negative real number, with values closer to zero being more harmonic.

Table 1 illustrates harmony calculations in MaxEnt using an example of free variation, English *t/d*-deletion, based on Coetzee & Pater (2011). This table shows three tableaux that compare faithful and deleted realizations of stem-final, postconsonantal [t] in three environments: (*a*) prepausal, (*b*) preconsonantal, and (*c*) prevocalic. There is one constraint that penalizes postconsonantal [t] (*Ct), one general Max constraint, and two contextual variants of Max, one specific to the prevocalic context (Max-P-V) and one to the phrase-final context (Max-Fin). The table shows the harmony calculations assuming weights of <4, 1, 2, 3> for the constraints <*Ct, Max-P-V, Max-Fin, Max>, respectively. Each violation has a numeric value of −1. In tableau *a*, Max-Fin (2)

Table 1 Example of English variable *t/d*-deletion[a]

	Input	Output	*Ct $w_1 = 4$	Max-P-V $w_2 = 1$	Max-Fin $w_3 = 2$	Max $w_4 = 3$	Harmony	Probability
a	/Ct/	[Ct]	–1				$(-1)^* w_1 = -4$	≅ 73.1%
		[C_]			–1	–1	$(-1)^* w_3 + (-1)^* w_4 = -5$	≅ 26.9%
b	/CtC/	[CtC]	–1				$(-1)^* w_1 = -4$	≅ 26.9%
		[C_C]				–1	$(-1)^* w_4 = -3$	≅ 73.1%
c	/CtV/	[CtV]	–1				$(-1)^* w_1 = -4$	= 50.0%
		[C_V]		–1		–1	$(-1)^* w_2 + (-1)^* w_4 = -4$	= 50.0%

[a]Table based on Coetzee & Pater (2011).

and MAX (3) together assign a harmony of −5 to the deletion candidate (/Ct/, [C_]), whereas the faithful candidate (/Ct/, [Ct]) violates only *CT, receiving a harmony of −4. Thus, in the prepausal context, the faithful candidate has higher harmony and is preferred according to these weights. In the second competition representing the preconsonantal context (tableau *b*), the deletion candidate (/CtC/, [C_C]) violates only the general MAX (3), making it more harmonic than the faithful candidate (/CtC/, [CtC]). In tableau *c*, representing the prevocalic context, the two candidates tie.

In MaxEnt, harmony is used to define the conditional probability $P(y|x)$ of an output y given an input x:

$$P(y|x) = \frac{\exp\left(\sum_{c\in C} w_c v_c(x,y)\right)}{Z}.$$

The probability is proportional to the exponential of the harmony, and the constant Z is a normalizing term to ensure the conditional probabilities sum to 1 for each input. Specifically, Z is the sum of the exponentiated harmonies for all output candidates $y \in Y(x)$ for a given input x:

$$Z = \sum_{y\in Y(x)} \exp\left(\sum_{c\in C} w_c v_c(x,y)\right).$$

The last column of **Table 1** shows the MaxEnt probabilities for each tableau. In tableau *a*, the faithful candidate has probability $[\exp(-4)]/[\exp(-4)+\exp(-5)] \cong 73.1\%$, whereas the deletion candidate has probability $[\exp(-5)]/[\exp(-4)+\exp(-5)] \cong 26.9\%$. With these weights, the probabilities of the faithful candidates in tableaux *b* and *c* are roughly 26.9% and 50%, respectively.

Coetzee & Pater (2011) show how different weightings of these constraints can account for empirically observed, phonologically conditioned rates of *t/d*-deletion across a wide range of English dialects. MaxEnt, Stochastic OT, and Noisy HG are all able to achieve a close fit with the observed rates. Beyond *t/d*-deletion, there are many other successful examples of modeling free variation in the literature using these frameworks (see, e.g., Boersma & Hayes 2001, Coetzee & Pater 2008, Goldwater & Johnson 2003).

Online learning algorithm: an algorithm that incrementally processes learning data, making updates on a word-by-word basis

Batch learning algorithm: an algorithm that processes the learning data en masse, making updates after consulting the entire data set

Error-driven learning: a learning strategy that assumes updates to learners' hypotheses occur when their current hypothesis fails to generate a match with the observed data

2.2. Learning Free Variation

Numerous successful algorithms have been developed for learning categorical OT rankings and HG weightings from full structural descriptions (Boersma & Pater 2016, Goldwater & Johnson 2003, Jäger 2007, Magri 2012, Soderstrom et al. 2006, Tesar 1995). Learning from full structural descriptions means that the learner is provided with access to all representations referenced by constraints, including hidden representations. How models have been extended to move beyond this simplifying assumption is the focus of the next section.

Given full structural descriptions, a number of algorithms exist for both OT and HG which are guaranteed to find a categorical target grammar for any set of input–output pairs, as long as such a target grammar exists. Both online learning algorithms, which process learning data one by one, and batch learning algorithms, which compute updates after consulting all the data, have been developed. The online error-driven constraint demotion (EDCD) algorithm (Tesar 1995) has been particularly influential and forms the basis of a number of the frequency-sensitive models discussed below. "Error-driven" (Gibson & Wexler 1994, Rosenblatt 1958, Wexler & Culicover 1980) means that updates to the grammar are triggered when the learner's own predicted output fails to match the observed output in the learning data.

Learning free variation presents a greater challenge since the learner must account not only for categorical properties of the data but also for quantitative patterns. Variable patterns introduce inconsistency, which means there is no categorical ranking or weighting that can account for all the

data. Instead, the learning task involves identifying a stochastic weighting or ranking that captures the rates of occurrence of observed phonological variants on the basis of input–output pairs, their relative frequencies, and the constraint violations of all candidates for each input. Sensitivity to frequency is inherent to the learning task: Only learning models with sensitivity to quantitative properties can capture the numerical tendencies in the data.

Likelihood maximization: an objective function for fitting parameters of generative statistical models that prefers hypotheses that assign maximal probability to the observed data, favoring hypotheses that tightly fit the observed distributions

Online algorithms for learning free variation include the error-driven Gradual Learning Algorithm for Stochastic OT (OT-GLA; Boersma 1997, Boersma & Hayes 2001) and the closely related version for Noisy HG (HG-GLA; Boersma & Pater 2016). Grammar updates work similarly in both algorithms. Suppose the learning data includes the faithful input–output pair (/Ct/, [Ct]) from **Table 1**, and the learner incorrectly selects the deleting (/Ct/, [C_]) as the winning candidate—an error. The learner compares the constraint violations of the observed candidate (/Ct/, [Ct]) with the violations of the error (/Ct/, [C_]), slightly demoting constraints that favor the error (*CT) and slightly promoting constraints that favor the observed form (MAX and MAX-FIN). Under the assumption that the current weights are <4, 1, 2, 3> and the learning rate (how much weights are adjusted on each update) is 0.1, the updated weights will be <3.9, 1, 2.1, 3.1>, increasing the probability of the faithful candidate to approximately 78.6% (from 73.1%). In general, each update results in a small adjustment to the probability distribution defined by the stochastic grammar, making the observed candidate slightly more likely than the error (for technical details, see Boersma & Pater 2016, Jarosz 2016a).

When there is free variation, the same input occurs with multiple different outputs in the learning data. For example, the word *cost* in English might sometimes be realized as [kɑs] and sometimes as [kɑst] in the same environment. This creates inconsistency, but the GLA is oblivious to this, updating the grammar slightly on the basis of each observation independently. Each time the learner observes (/kɑst/, [kɑst]), it must predict [kɑst] as the output, and [kɑs] will be treated as an error, whereas each time the learner observes (/kɑst/, [kɑs]), the opposite is true. The right outcome in each case is unpredictable, so the learner will continue to make small updates in opposite directions throughout learning, but these updates will be made in proportion to the rate at which these variants occur in the data. Updates favoring the more frequent variant will be made more often, and the learned grammar will therefore generate the frequent variant more often. The model's sensitivity to input frequency is essential for learning free variation and matching the relative rates of occurrence in the data. When exposed to systematic free variation, the GLA yields variable final grammars which generally match the empirical rates of variation quite well. The GLA often works well in practice; however, it is not guaranteed to find a grammar compatible with the data in all cases, even for categorical patterns (Pater 2008).

MaxEnt models have been widely utilized in psycholinguistics and sociolinguistics, as well as outside linguistics in a variety of machine learning and natural language processing contexts. Indeed, MaxEnt is simply another name for multinomial logistic regression, which is one of the most broadly applied and widely understood statistical models in all of the social sciences. There are numerous well-understood optimization algorithms for finding weights that optimize fit with the data (Berger et al. 1996, Della Pietra et al. 1997, Goldwater & Johnson 2003, Hayes & Wilson 2008, Jäger 2007, Johnson 2002, Wilson 2006). For example, standard algorithms exist for performing (stochastic) gradient descent (SGD) for these models, and they are guaranteed to find the weights that best fit the observed distribution. Online learning for MaxEnt is a specific application of SGD, and Jäger (2007) shows that SGD updates for MaxEnt look exactly like the HG-GLA updates. Data fit in MaxEnt modeling is usually defined in terms of likelihood maximization: Likelihood is maximized when the learned grammar matches observed frequencies in the data as well as possible. In MaxEnt models it is also straightforward to include priors, or regularization terms, in the objective function to keep weights low and prevent overfitting (Goldwater & Johnson 2003) or

to encode other biases on weightings of constraints (Pater et al. 2012, Wilson 2006). This capacity plays an important role in modeling the learning biases discussed in Section 4.

For the GLA, frequency sensitivity is inherent to the mechanics of learning, while for MaxEnt, frequency sensitivity is part of the learner's objective function. In both cases, the statistical nature of the model is essential for learning and representing free variation.

2.3. Gradient Phonotactics

The MaxEnt, Stochastic OT, and Noisy HG frameworks can also be used for modeling graded acceptability and phonotactics (Boersma & Hayes 2001, Coetzee & Pater 2008, Hayes & Wilson 2008).[2] The most common approach follows Hayes & Wilson (2008) in using only markedness constraints to define a probability distribution over the entire space of possible word forms in the language. Rather than defining probabilistic mappings (conditional distributions over outputs for each input), phonotactic grammars simply define a single distribution over all possible output forms. Each possible word form has an associated probability relative to other possible word forms, and this probability is determined by the weights of the markedness constraints that each form violates. In one application, Hayes and Wilson used a MaxEnt model to represent graded acceptability of English onset clusters. Highly weighted constraints, such as *ŋ, penalized onsets heavily, reducing their probability to near zero, while weaker constraints, such as *[+cont,−strid] (no interdental fricatives), reduced the probability slightly. The frequency sensitivity of the model is what allows it to represent weak penalties for observed yet statistically underrepresented patterns.

The predicted probabilities of various forms can be numerically transformed and correlated with acceptability scales or other behavioral measures. The Phonotactic Learner (Hayes & Wilson 2008) has been especially broadly applied in recent years and has performed well in predicting experimentally elicited phonotactic scales, numerical ratings based on human well-formedness judgments of nonce words (see, e.g., Albright 2009, Daland et al. 2011). For example, Hayes and Wilson showed that their MaxEnt grammar for English onsets, whose weights were fitted based on the type frequency of actual English onsets, correlates strongly with human ratings of nonce words. These applications are discussed further in Section 4. In addition to dealing with inconsistency, the Phonotactic Learner takes on a hidden structure learning problem, learning constraints, which is discussed further in Section 3.

2.4. Classes, Exceptions, and Lexicalized Variation

Learning of classes, exceptions, and lexicalized variation presents both inconsistency and hidden structure challenges: Phonologically similar morphemes behave differently in the same environments, and the learner must infer the hidden classification underlying this inconsistency. If the learner is faced with even a few exceptions to a general pattern, they must infer which examples should be treated as exceptions and which can be treated as part of the general pattern. Similarly, if the learning data have lexical strata with distinct phonological properties, the learner must infer which examples fall into each stratum while learning the grammars corresponding to these strata and how they differ from one another.

Due to the difficulty of this learning task, most approaches are rather recent. The earliest research on learning lexical exceptionality in a constraint-based framework (Becker 2009, Coetzee

[2]For other approaches to modeling gradient phonotactics, see Albright (2009), Bailey & Hahn (2001), Coleman & Pierrehumbert (1997), Frisch et al. (2000), and Vitevitch & Luce (2004).

2009, Pater 2010) builds on the categorical constraint learning algorithm Recursive Constraint Demotion (RCD) and its ability to detect inconsistency (Tesar & Smolensky 1998). RCD keeps track of winner–loser pairs, efficiently finds a ranking that favors all winners over losers, if one exists, and efficiently detects inconsistency otherwise. When there are exceptions in the data, there will be inconsistency. Pater (2010) proposes an extension of this algorithm that constructs lexically specific constraints for the data forms that triggered the inconsistency. These lexically specific constraints are indexed to the deviant morphemes and can be ranked separately from their general versions to resolve the inconsistency.

Winner–loser pairs: in constraint-based learning, a pair of candidates, one of which is the observed form (winner) and the other a competitor (loser), together with their constraint violations; error-driven learning can be used to identify informative losers for each winner

Subset problem: the challenge of learning a restrictive grammar that captures systematic prohibitions and regularities in the language without overgeneralizing on unseen data

While the RCD-based exceptionality approach can deal with one kind of inconsistency (exceptions), it cannot cope with learning data that have exceptions and other kinds of inconsistency or ambiguity, such as variability or hidden structure. Various approaches to learning exceptions or classes in the face of variability have recently been developed by extending frequency-sensitive approaches such as those discussed in the previous two subsections (Nazarov 2016, 2018; Pater et al. 2012; Shih 2018). Although the details vary, all of these approaches crucially rely on the ability to model general statistical trends in the learning data while allowing individual lexical items the ability to counter the broader language-wide grammatical pressures. Related modeling research focuses on the gradient productivity of morphophonological transformations using rules (Albright & Hayes 2003) and constraints (Allen & Becker 2015, Becker & Gouskova 2016, Moore-Cantwell & Staubs 2014).

A related empirical and theoretical problem of particular interest in recent modeling research is that of gradient, or patterned, exceptionality. This type of exceptionality provides a particularly powerful empirical argument for quantitative models of phonological knowledge because existing psycholinguistic findings indicate that language learners extract statistical generalizations about phonological alternations even when alternations are lexicalized. Numerous experimental studies across multiple languages have shown that modeling speakers' generalization abilities requires the capacity to predict the fixed behavior of particular lexical items while simultaneously making gradient predictions for novel forms (Becker et al. 2011, Ernestus & Baayen 2003, Gouskova & Becker 2013, Hayes & Londe 2006, Zuraw 2000). For example, Zuraw (2000) shows that, across the lexicon in Tagalog, the rate of nasal substitution is statistically conditioned by phonological factors—voicing and place—and that native speakers reproduce these statistical trends for nonce words even though most prefix–stem combinations exhibit fixed behavior. Approaches to this problem model the lawful, phonologically conditioned statistical patterns in the lexicon using the GLA or MaxEnt model while incorporating constraints that allow individual lexical items' memorized pronunciations to be utilized when available (Moore-Cantwell & Pater 2016, Smith 2015, Zuraw 2000).

Lexicalized variation presents a version of the notoriously difficult subset problem (Berwick 1985). Since the target grammar requires lexicalization, and lexicalization perfectly accounts for the learning data, what prevents the learner from simply memorizing the exceptions and failing to learn anything general about the language-wide phonological patterns and restrictions? Put differently, what ensures that the learner will acquire a grammar that generalizes appropriately beyond the learning data? Several recent studies have shown that statistical models such as MaxEnt and the GLA for Stochastic OT generally learn language-wide patterns more quickly than they do lexically specific patterns (Moore-Cantwell & Pater 2016, Pater et al. 2012, Zuraw 2000). This capacity is, again, dependent on these models' sensitivity to frequency: All the learning data support language-wide patterns, while support for lexically specific patterns occurs rarely, only when a particular lexical item is observed. In one case study, Moore-Cantwell & Pater (2016) showed that a MaxEnt model learned default stress patterns more robustly when exceptions were rarer in the data. In general, statistically sensitive models naturally favor learning of

language-wide preferences more quickly or more robustly and idiosyncratic properties of lexical items more slowly or less robustly because statistical learning is sensitive to the quantity of data supporting each generalization.

3. LEARNING HIDDEN PHONOLOGICAL STRUCTURE

Learning of hidden phonological structure pushes the bounds of current learnability capabilities. In the presence of hidden structure, no known approach is guaranteed to succeed at efficiently learning every (arbitrary) phonological system. To deal with the massive ambiguity created by hidden structure, models place restrictions on the kinds of phonological patterns that can be learned in principle or learned reliably well. For frequency-sensitive models, it also means that quantitative properties of the data can dramatically influence learning outcomes. In either case, certain patterns or phenomena are predicted to be more difficult (or impossible) to learn. Modeling thus raises difficult and important questions about the kinds of patterns and representations learning models must account for and the kinds of biases that are needed. What are the limits on learnability, and to what extent are observable typological generalizations derivable from these limits? Modeling hidden structure learning also affords a unique opportunity to investigate the richness and universality of phonological representations. Which aspects of phonological representations must be innate, and which can be acquired? How abstract and structured is phonological knowledge? Which theoretical frameworks and assumptions lead to better learning outcomes or better fits to behavioral observations? Answering these questions requires a tight connection between computational modeling and the empirical sources of evidence for learning outcomes and learning biases: typology, psycholinguistic studies, and sound change.

This section does not attempt a comprehensive review of the rich and ever-growing literature on hidden structure learning in phonology (for recent overviews, see Jarosz 2013, 2015, 2016a; Tesar 2013). Rather, after highlighting some of the unique challenges posed by hidden structure and the developments that led to the existing range of solutions, the section outlines some of the major learnability results and discusses novel insights into long-standing debates that recent modeling research has begun to produce.

3.1. Hidden Structure Challenges

Ambiguity is particularly challenging for hidden structure learning: The space of possible analyses the learner must be capable of navigating is too large to search exhaustively. Even when the space is finite, such as with metrical footing, it grows exponentially, or worse, with the number of words, features, or constraints (Prince 2010). In the case of learning abstract URs, rules, or constraints, the space is potentially infinite, even for categorical languages. To take a simple example, in a language that deletes final consonants, there is no bound in principle on the number of final consonants that may be posited underlyingly. Likewise, there is no bound in principle on the maximal length of phonotactic constraints (see Hayes & Wilson 2008) or on the length of phonological contexts of rules (see Albright & Hayes 2003). Therefore, the learner must somehow constrain their search through this vast space of possibilities while finding ways to explain generalizations that can only be discovered with reference to patterns across many lexical items.

One kind of ambiguity that arises in hidden structure learning is the credit (or blame) problem (Dresher 1999), which has a "chicken-and-egg" character. When the learner's current hypothesis makes an erroneous prediction, hidden structure prevents the learner from directly observing the source of the error. For example, when simultaneously learning phonological mappings and URs, an error could be the result of an incorrect lexical representation or an incorrect phonological

mapping, and the learner must somehow determine which should be blamed. Similarly, since metrical footing cannot be directly observed, when an error occurs, it is not clear which constraints, parameters, or rules are to blame. For example, when the learner observes a trisyllabic word with stress on the medial syllable, such as [tɛˈlɛfɔn], it is not clear whether this form supports left-aligned iambs [(tɛˈlɛ)fɔn] or right-aligned trochees [tɛ(ˈlɛfɔn)]. If the learner knew the target footing, they could determine the constraint violations of the observed form and the required update to the grammar using one of the algorithms discussed in Section 2.1. Conversely, if they knew the target grammar, they could make inferences about the footing of this form. Since learners have prior knowledge of neither, they must overcome this chicken-and-egg ambiguity if they are to make any progress with hidden structure learning.

Bias–variance trade-off: the balance between tightly fitting observed data (low bias) and generalizing appropriately to unseen data (low variance)

Another source of ambiguity in hidden structure learning is the relative breadth or narrowness of inferred generalizations. The subset problem discussed above for exceptionality arises when learning URs or any other lexically specific properties.[3] A related issue arises when learning rules or constraints: How broad or narrow should constraints or rules be? The learner must generalize from the incomplete data sample representing the target language's patterns. The observed data (and, indeed, entire language lexicons) do not contain every combination of segments, features, and contexts to which a rule or constraint is potentially relevant (for related discussion, see Wilson & Gallagher 2018). On what basis does the learner generalize, and how broadly? Relatedly, when does the learner have enough evidence to abstract a general rule or constraint rather than treating a pattern as accidental? Modeling human learning requires just the right balance between restrictively fitting the observed data—with its noise and accidental gaps—and generalizing appropriately to "similar" unseen data. This is sometimes called the bias–variance trade-off. Defining precisely what "similar" means in phonological learning—and how features, representations, and substantive and quantitative factors influence this process—is an important area of ongoing research.

3.2. Approaches and Progress

A common theme unifies many of the results summarized in this section: Much of the progress on hidden structure learning in phonology can be traced to a productive integration of linguistic theory with machine learning approaches and statistical methods used throughout the social sciences. Numerous models discussed in this section build on well-studied techniques such as likelihood maximization for incomplete data (Dempster et al. 1977), minimum description length (Solomonoff 1964), and information theory and maximum entropy modeling (Berger et al. 1996). These successes are a testament to the possibilities that actively integrative computational modeling research can yield.

The previous section argued that frequency-sensitive learning models are necessary for modeling human learning of quantitative patterns such as variability, gradient well-formedness, and exceptionality. Frequency-sensitive learning approaches can also provide a way to "break into" the chicken-and-egg ambiguity that hidden structure creates. Modeling research on various linguistic interfaces has shown that learning of quantitative preferences, even if those preferences are based on incomplete or noisy data, can guide subsequent learning. For example, learning of phonotactic distributions can facilitate learning of phonological rules (Calamaro & Jarosz 2015, Le Calvez et al. 2007, Peperkamp et al. 2006) and word boundaries (Blanchard et al. 2010, Daland &

[3]There is a sizable literature on strategies that favor restrictive phonological grammars (Alderete & Tesar 2002; Hayes 2004; Jarosz 2006, 2009; Jesney & Tessier 2011; Prince & Tesar 2004; Tesar & Prince 2007; Tessier 2009).

Pierrehumbert 2011, Jarosz & Johnson 2013, Johnson 2008a) from noisy corpus data. Learning of lexical entries (Feldman et al. 2009) and phonemes (Dillon et al. 2013) can help with the learning of phonetic categories, and simultaneous learning of word co-occurrences and word boundaries can be mutually informing (Goldwater et al. 2009, Johnson 2008b). Quantitative modeling also enables general and principled solutions to the subset problem, making it possible to formalize mathematically how learners balance conflicting considerations such as the simultaneous pressures to tightly fit ambiguous and gappy observed data and to extract broad and simple generalizations (Dillon et al. 2013, Hayes & Wilson 2008, Jarosz 2006, Rasin & Katzir 2016, Wilson & Gallagher 2018).

3.3. Overview of Modeling Results

This subsection reviews developments on a variety of hidden structure learning problems in phonology, emphasizing the contributions of cross-disciplinary research to progress in these areas and highlighting the insights that the resulting models have revealed.

3.3.1. Prosodic structure. One of the best-studied hidden structure learning problems in phonology is that of metrical structure. While metrical structure has attracted particular attention, many of the approaches discussed below could be applied equally well to other types of abstract representations, such as syllables or autosegments.

Modeling learning of metrical parameter settings in the Principles and Parameters framework (Chomsky 1981) provides a concrete example of how computational modeling can contribute to debates regarding fundamental questions about innate linguistic knowledge. To address the overwhelming ambiguity created by metrical footing, pioneering research (Dresher 1999, Dresher & Kaye 1990) developed an approach called cue-based learning. In the cue-based learning approach, each parameter is innately associated with a "cue"—a pattern in the data that prompts the learner to set that parameter to a certain value. For example, upon observing that stress occasionally falls on the rightmost syllable, the learner may determine that (right) extrametricality is set to "off" in the target language. Dresher and Kaye hypothesized that, in addition to innate cues, successful learning requires that parameters have default settings and an inherent ordering. More recently, Pearl (2011) applied a statistical learning model proposed for syntactic parameters (Yang 2002) to the learning of metrical structure, which made it possible to learn parameter settings from noisy data. In support of innate language-learning processes, Pearl found that the statistical learning algorithm needed to be supplemented with cues and parameter ordering. However, building on statistical machine learning approaches (see Jarosz 2015), Nazarov & Jarosz (2017) recently found that the more nuanced statistical inference capabilities of their proposed learning model, the Expectation Driven Parameter Learner, allowed it to succeed at learning a wide range of metrical parameter systems without the need for cues, default settings, or inherent ordering, thereby weakening the arguments for innate domain-specific learning processes.

Prosodic structure was also the first hidden structure domain addressed in OT. Tesar & Smolensky (1998, 2000) proposed a parsing strategy called Robust Interpretive Parsing (RIP) that allows the learner to make an educated guess about the prosodic structure of the learning data. RIP adapts a standard statistical machine learning approach called Expectation Maximization (EM) to the categorical OT setting (Dempster et al. 1977). The basic intuition behind RIP (and EM) is that the learner can use their own current grammar to choose among competing interpretations, or parses, of the overt forms in the data. This allows the learner to circumvent the chicken-and-egg problem discussed above: They use their current grammar to guess the hidden

structure in the learning data, and then they use that hidden structure to calculate the update to their grammar. Returning to the example of structurally ambiguous [tɛˈlɛfɔn], RIP works by limiting the candidate set to metrical parses of the observed form (e.g., [(tɛˈlɛ)fɔn] and [tɛ(ˈlɛfɔn)]) and selecting whichever parse is optimal according to the current ranking. The candidate corresponding to that fully structured form is then compared with the learner's own production, which is the optimal candidate among all possible stress assignments for /tɛlɛfɔn/ under the current grammar. If there is a mismatch, the constraint ranking is updated as usual on the basis of the constraint violations of both candidates.

The parsing strategy proposed by Tesar & Smolensky (1998, 2000) was later extended to the stochastic setting, where it has been used to explore learning biases and compare the learning consequences of weighted versus ranked constraints (Apoussidou 2007, Apoussidou & Boersma 2003, Boersma 2003, Boersma & Pater 2016, Breteler 2018, Jarosz 2013). Boersma (2003) extended this approach to the OT-GLA, while Boersma & Pater (2016) extended it to the HG-GLA and presented simulations comparing the performance of RIP as applied to categorical OT, OT-GLA, and HG-GLA. They found that the statistical models and especially those with weighted constraints performed best, suggesting a potential learnability advantage of HG over OT. Jarosz (2013) subsequently showed that the original formulation of RIP for the GLA selects parses of overt data in a way that is incompatible with the learner's production grammar, leading to internal inconsistencies and inefficiencies during learning. Jarosz (2013) proposed two alternative parsing strategies that incorporated insights from statistical machine learning to enhance the learner's utilization of their probabilistic knowledge during parsing. The essential insight behind the alternative parsing strategies, Resampling Robust Interpretive Parsing (RRIP) and Expected Interpretive Parsing (EIP), again comes from EM: Parses of overt forms should be sampled in proportion to their probability according to the production grammar. Jarosz showed these strategies substantially outperformed RIP for both OT and HG. Of greater theoretical interest is the associated finding that the improvements affected OT more than HG, leveling the performance of the OT and HG learning models and thus revealing that the OT disadvantage discovered by Boersma & Pater (2016) was due to properties specific to RIP rather than OT per se. In follow-up research, Jarosz (2015) drew further inspiration from EM and proposed a novel learning approach for probabilistic OT, whose performance on learning metrical structure slightly surpasses the best parsing strategies and extends to other kinds of hidden structure, such as lexical representations and derivations (discussed in the next subsection).

In summary, insights from statistical machine learning have been crucial to several key advances in the learning of hidden prosodic structure. Not only have these advances improved the performance and robustness of phonological learning models, they have also been integral to debates about the contents of the language acquisition device more generally.

3.3.2. Lexical representations. Much of the earliest research on learning URs focused on lexical accent. Even when learning is restricted to learning underlying features of observed segments—that is, when insertion and deletion mappings are not considered—the space of possible URs is exponentially large. To be computationally feasible, models must either restrict the feature values or forms considered by the learner to a limited set observed on the surface (Hayes 2004, Pater et al. 2012, Tesar 2006) or find computational strategies for efficiently searching through a larger space of abstract URs (Jarosz 2015, Merchant 2008, Tesar 2013). Various representational approaches have been developed for modeling lexical properties. Some assume the traditional UR that the grammar uses as the input to the phonological mapping (Akers 2012, Dresher 2016, Jarosz 2015, Merchant 2008, Tesar 2013), while others rely on lexical (or UR) constraints that interact with grammatical constraints in parallel (Apoussidou 2007, Pater et al. 2012). With the

latter approach, the models developed for learning of structural ambiguity (e.g., RIP, RRIP, EIP) can also be applied to the learning of lexical representations.[4]

The computational pressure is intensified when alternations involving insertion and deletion are considered (Alderete & Tesar 2002; Cotterell et al. 2015; Jarosz 2006, 2009; Merchant 2008; O'Hara 2017; Pater et al. 2012; Rasin & Katzir 2016). As discussed above, learning of deletion mappings opens the door to a potentially infinite space of abstract URs. To model this aspect of phonological learning, assumptions about the range of lexical options available to the learner must be made explicit. Research on learning of such alternations thus necessarily makes claims about the abstractness or concreteness of lexical representations and the restrictions on possible types of alternations, reviving classic debates on abstractness in the phonological literature (Kisseberth & Kenstowicz 1977). Currently, these limits are not well understood; however, modeling research is beginning to provide new arguments for both abstract (O'Hara 2017) and concrete lexical representations (Allen & Becker 2015). For instance, O'Hara (2017) shows that a MaxEnt learner equipped with the capacity to consider a range of abstract URs nevertheless inherently favors less-abstract lexical entries, deriving principles of economy proposed in the phonological literature. There is great potential to make explicit the trade-offs between more abstract lexical representations and the ability to (efficiently) learn attested kinds of alternations by applying, extending, and testing the current range of learning models.

3.3.3. Derivations and intermediate representations. Learning of serial derivations—phonological mappings that allow intermediate representations between input and output—is probably the least well understood learning problem in phonology. Most of the progress on this task has occurred in the last several years, building on machine learning techniques and solutions developed for other hidden structure problems.

Prior to OT, there was limited research on learning of rules and rule ordering, and even learning of individual rules (let alone a system of ordered rules) given pairs of underlying and surface forms continues to be a challenging problem. Johnson (1984) proposed a procedure for learning of URs and ordered rules from paradigmatic information, but this procedure makes strong simplifying assumptions about the types of rules and interactions allowed—for example, insertion and deletion are not considered. Gildea & Jurafsky (1996) showed that learning a single simple rule from naturalistic data, English flapping, presents numerous challenges. Learning is unsuccessful even though the algorithm makes strong restrictions on possible mappings (they must be subsequential; Mohri 1997) and is guaranteed to learn the target mapping in the limit (Oncina et al. 1993). Gildea and Jurafsky showed that the problem arises due to lack of sufficient restrictions on generalization. As discussed above, naturalistic data do not provide every combination of features or segments that instantiate a rule or pattern. Without biases favoring more natural[5] phonological rules, the algorithm fails to generalize appropriately to unseen data. More recent studies on subregular formalizations of phonology have investigated even tighter formal restrictions on permissible mappings (Chandlee et al. 2014, Chandlee & Heinz 2018). However, these learning procedures still assume that the learner observes input–output pairs and all combinations of segments that instantiate the pattern.

Frequency-sensitive approaches have also been developed recently. Rasin et al. (2015) pursue an approach using principles of minimum description length (Solomonoff 1964) to learn both URs

[4]See Jarosz (2015) for a discussion of why RIP cannot be applied to learning of traditional URs.

[5]Gildea & Jurafsky (1996) proposed three biases that improved learning outcomes: faithfulness, community, and context. See Calamaro & Jarosz (2015) and Peperkamp et al. (2006) for related research on biases needed for learning rules from noisy corpus data.

and ordered rules. Staubs & Pater (2016) and Jarosz (2016b) propose novel approaches for learning serial derivations in Harmonic Serialism (HS; McCarthy 2000, Prince & Smolensky 2004), while Nazarov & Pater (2017) model learning of derivations in a MaxEnt version of the Stratal OT framework (Bermúdez-Otero 1999, Kiparsky 2000). These approaches have the potential to address long-standing conjectures about the naturalness of process interactions (Kiparsky 1968, 1971). Indeed, initial simulation results are starting to provide evidence that learnability may be able to capture Kiparsky's hypothesized biases under certain conditions (Jarosz 2016b, Nazarov & Pater 2017). Both the HS (Jarosz 2016b) and Stratal MaxEnt (Nazarov & Pater 2017) models predict easier learning of certain transparent process interactions over opaque interactions (Kiparsky 1971), and under certain conditions, the HS model (Jarosz 2016b) also predicts easier learning of feeding and counterbleeding interactions over bleeding and counterfeeding interactions (Kiparsky 1968). Psycholinguistic results have provided initial support for these biases as well, suggesting that human learning of artificial languages in the lab may indeed be sensitive to similar pressures (Prickett 2018a).

These findings suggest important questions that require further empirical investigation, such as the relative rates and availability of phonological contexts in which processes interact and occur independently. A complete understanding of how learning biases may influence diachronic rule reordering and rule loss will require substantial further computational, experimental, and typological work. Investigating how learning of URs affects these pressures is a concrete area for future research.

3.3.4. Constraints. A decade ago, Hayes & Wilson (2008) introduced a MaxEnt model and an associated software package for learning of phonological constraints from natural language data that has had a transformative impact on the field. Previously, most constraint-based learning models made the traditional OT assumption that constraints are innate and therefore provided to the learner at the outset. Hayes and Wilson demonstrated, however, that many phonological generalizations can be successfully induced from naturalistic data by constructing constraints that account for underattested patterns. Crucially, they also showed that successful learning required reference to abstract phonological representations: features, natural classes, and autosegmental tiers. This work inspired a substantial body of follow-up research, discussed in the next section, investigating computationally and experimentally what biases are required to account for human learning and generalization.

To formalize underattestation and learn restrictive phonotactic grammars, Hayes and Wilson found an efficient solution to a difficult computational problem. To calculate weight updates in MaxEnt models, the learner must compare the number of observed violations of each constraint in the learning data with the expected number of violations of that constraint given the current grammar and weights. Calculation of the observed violations is straightforward: It involves summing the violations of each constraint in the observed data. However, the expected violations require estimating the number of violations that result from applying the current constraints to a base of all possible phonological forms—an infinite set, in principle. Concretely, to calculate weight updates and learn constraints for unattested patterns, the learner must have access to losing candidates, that is, unattested patterns. Only by noticing that a constraint such as *#ŋ correctly rules out unattested forms that would otherwise be predicted can the learner induce this constraint and weight it highly. Hayes and Wilson's solution is to represent the set of possible forms and their associated constraint violation vectors in a finite-state machine. To simplify computation, they restrict attention to the finite set of all possible forms no longer than those in the learning data. Given the finite-state representation, it is possible to efficiently sum the violations of each

constraint over all the forms in the machine (Eisner 2002)—exactly what is needed to estimate expected violations.

Comparing expected and observed distributions also provides a way to quantify the robustness of a phonological generalization to determine whether a pattern supports a general constraint or represents an accidental gap (Wilson & Gallagher 2018). As discussed above, accidental gaps are characteristic of natural language input and must be distinguished from robust restrictions. It is only through sensitivity to quantitative patterns that learners can make such crucial distinctions given gappy and noisy learning data.

Wug test: a task used to test productivity of morphophonological knowledge by asking speakers to produce (or rate) a morphological derivative of a nonce word

4. MODELING HUMAN LEARNING, GENERALIZATION, AND TYPOLOGY

So far, this article has argued that sensitivity to quantitative patterns in natural language data is essential for modeling variation, gradience, and exceptionality and for addressing learning challenges posed by hidden structure. This final section reviews some of the most significant findings on human learning and generalization that such models have revealed. Natural language contains statistical information, and learners are sensitive to this information—only by modeling learners' sensitivity to this information can we draw firm conclusions about what learners can and cannot infer from data. Frequency-sensitive models have shown that learners can successfully extract more from their language input than many imagined was possible. One example, discussed above, is Hayes & Wilson's (2008) MaxEnt model for inducing constraints and their weights from unlabeled data. Another is Nazarov & Jarosz's (2017) Expectation Driven Parameter Learner, which learns stress parameter settings without language-specific learning mechanisms. At the same time, the integration of modeling and behavioral research has provided concrete evidence of biases and restrictions on human learning and generalization that could help explain much about typology, language change and language development.

4.1. Modeling First Language Acquisition

The learning models discussed above have been applied to a range of behavioral tasks, each of which provides unique insights into the learning biases that shape first language acquisition. One way to study learning biases is to compare the predictions of models exposed to natural language data representative of learners' first language input to adults' behavior on linguistic tasks in their native language. To approximate learners' language input, models are typically provided with large data sets of phonetically or phonemically transcribed words or paradigms in the target language. By comparing predictions of models making different theoretical or representational assumptions, one can make inferences about the likely contents of the human language acquisition device. Wug tests (Berko 1958) are used to study speakers' productive knowledge of morphophonological alternations in their language and can be compared with models that generate predictions about alternations. Another way to probe speakers' knowledge is by comparing acceptability judgments on phonotactic patterns or alternations with models' numerical predictions about the relative goodness of various patterns. In both cases, models are tasked with predicting speakers' end-state knowledge of their native language phonologies, which makes it possible to directly investigate biases that affect the outcomes of first language acquisition.

This approach has produced a sizable literature leading to discoveries about a wide range of learning biases needed to successfully model phonological acquisition. Initial research using quantitative models demonstrated the success of stochastic grammars capable of extracting abstract generalizations from the lexicon (Albright & Hayes 2003, Boersma & Hayes 2001, Coleman

Substantive bias: a type of inductive, or analytic, bias that favors the learning of patterns with perceptual or articulatory motivations

Analytic bias: a cognitive predisposition, or inductive bias, that makes learners more receptive to some patterns than others

Channel bias: phonetically systematic errors in language transmission between speaker and hearer

& Pierrehumbert 1997). For example, Albright & Hayes (2003) showed that abstract rules better capture learning of morphophonological alternations than do analogical models that directly compute overall similarity with the lexicon. A subsequent series of studies demonstrated that sensitivity to abstract phonological representations, such as features, natural classes, syllables, and tiers, is needed to capture behavioral results (Albright 2009, Coetzee & Pater 2008, Daland et al. 2011, Hayes 2011, Hayes & Wilson 2008). For example, Daland et al. (2011) showed that several models can predict English speakers' acceptability ratings on nonce words with initial consonant clusters varying in their sonority profiles. Crucially, only models with the capacity to represent aspects of syllable structure and featural similarity can successfully predict speakers' gradient preferences for consonant clusters with higher sonority rises (Clements 1990, Selkirk 1982).

Perhaps the most broadly investigated question in recent modeling research concerns the role of phonetic naturalness and substantive bias (Becker et al. 2011; Berent et al. 2007; Davidson 2006; Hayes & Londe 2006; Hayes & White 2013; Hayes et al. 2009; Jarosz & Rysling 2017; O'Hara 2018; Prickett 2018b,c). It has long been observed that phonetic naturalness plays a role in shaping typology, yet the exact nature of this pressure continues to be a matter of debate. Are effects of naturalness encoded as hard grammatical universals in Universal Grammar (UG; Prince & Smolensky 2004) or as soft analytic biases in the language acquisition device (Hayes 1999, Moreton 2008, Wilson 2006), or do they affect language change indirectly via channel bias (Blevins 2004, Ohala 1993)? Although more research is needed, the emerging view that recent modeling work supports is that phonetic substance likely affects how easily or robustly patterns are learned, but it does not place categorical limits on learnability. For example, Jarosz & Rysling (2017) found that modeling Polish adults' phonotactic judgments on initial clusters with varying sonority profiles supported a combined role of frequency sensitivity and sensitivity to a soft substantive universal favoring larger sonority rises. While this may seem like an obvious conclusion, it conflicts with a prevailing view in the field that there are categorical, substantive constraints on possible, and therefore learnable, languages. One promising way to formalize soft inductive biases is via priors in MaxEnt, which can be used to incorporate phonetic difficulty (as formalized in, e.g., Steriade 2008), making it harder to learn high weights for phonetically unmotivated constraints (White 2017, Wilson 2006). However, much work remains to be done in formalizing exactly how substantive factors influence learning and understanding whether these kinds of pressures could give rise to universal generalizations observed crosslinguistically.

In summary, modeling of first language acquisition has provided evidence for the role of abstract phonological representations and soft substantive biases.[6]

4.2. Modeling Artificial Language Learning

Another approach that has been used to investigate learning biases is artificial language learning (ALL). In ALL, participants are presented with miniature languages in the lab and tested on their learning and generalization of those patterns. ALL differs from the first language acquisition process in numerous ways; however, it allows for precise control of the linguistic input that makes it possible to investigate the learning of patterns that cannot be easily found or manipulated in natural languages. In the ALL context, evidence for substantive bias (Finley & Badecker 2009, White 2017, Wilson 2006) has been rather weak and mixed (for a review, see Moreton & Pater 2012a). Since evidence from first language acquisition has demonstrated sensitivity to phonetic

[6]Modeling of the first language acquisition process in children has also supported a role for representational and substantive learning pressures (Boersma & Levelt 2000; Hayes 2004; Jarosz 2006, 2010, 2017; Jarosz et al. 2017; Jesney & Tessier 2011; Prince & Tesar 2004).

naturalness, this discrepancy is likely due to the differences between first language acquisition and ALL. One substantial difference is that first language learners must discover the phonetic categories of their first language and cope with perceptual and articulatory difficulties in acquiring them and the phonological system, whereas participants in ALL studies have already learned the categories and their relationships in their first language.

Nonetheless, ALL studies have yielded consistent evidence of another important learning bias: complexity. In general, patterns that require fewer features to express are easier for participants to learn (for a recent review, see Moreton & Pater 2012b). Moreton & Pater (2012a) show that it is important to keep the effect of complexity in mind when examining other learning pressures since complexity and naturalness are often correlated (see also Prickett 2018c). Formalizing simplicity and comparing its effects in linguistic and nonlinguistic domains have also been investigated (Moreton et al. 2015).

4.3. Modeling Diachrony and Typology

A growing body of research is using quantitative models of phonological learning to investigate soft learning biases that could be responsible for crosslinguistic tendencies and universals.[7] A standard assumption in OT is that the universal set of constraints should define the space of possible languages via factorial typology. Under this view, systematic gaps in the typology must be categorically ruled out by UG. This perspective precludes the possibility of modeling crosslinguistic tendencies rather than strict universals and overlooks pressures aside from UG that may be involved in shaping the observed typology, such as domain-general learning biases. As discussed above, models of phonological learning make predictions about the relative ease of learning of various patterns: Some patterns are learned more quickly or require fewer data than others, and when there is hidden structure, current models predict that some patterns should not be learned at all, at least not under all conditions.

Examining the correspondence between models' learning difficulties and typology has revealed a number of possible ways that learning might shape typology. Learning biases favoring certain process interactions over others are discussed in Section 3.3.3. Several recent studies have investigated how biases inherent to statistical learning may in part shape the typology of stress and tone systems (Breteler 2018, Stanton 2016, Staubs 2014). For example, Stanton (2016) shows that learning pressures provide a possible explanation for the absence of a stress pattern known as the midpoint pathology. The evidence necessary to distinguish this pattern from competing analyses of the same data occurs rarely in distributions representative of natural languages. Using the iterated learning paradigm (Kirby et al. 2004), Hughto (2018) shows, by simulating generations of child–parent learning interactions, that MaxEnt learning models can, over time, introduce biases into the typology that favor phonological systems which minimize free variation and cumulativity. Using interactive learning, Pater (2012) shows that preferences for systemic simplicity—wherein a language expresses a general preference across multiple contexts, such as uniform headedness across multiple categories—naturally emerge in MaxEnt learning models. Thus, modeling studies are beginning to provide evidence that learners' sensitivity to the distributional information in the language input may help explain some crosslinguistic tendencies. These results have broad implications for linguistic theory because they show that biases inherent to statistical learning can systematically skew the typological predictions derived from theoretical assumptions.

Iterated learning: a type of agent-based model that simulates vertical transmission of language across generations by modeling "parent–child" interactions in which one agent (the "parent") provides input from a target language to the other agent (the "child"), who eventually becomes the parent in the next generation

Cumulativity: a type of constraint interaction possible in weighted grammars wherein violations on lower-weighted constraints combine to overpower the preferences of higher-weighted constraints

Interactive learning: a type of agent-based model that simulates interactions between speakers within a generation to understand how communicative pressures may cause languages to drift over time

[7] Research on formal language characterizations of phonological patterns and processes provides a complementary perspective on typological restrictions (for an overview, see Heinz 2018). So far, there has been little work integrating formal language constraints on typology with the kind of quantitative modeling and abstract phonological representations that this article has argued are essential to modeling human learning in the face of noise and ambiguity (but see Lamont 2018, Yu 2017).

5. CONCLUSIONS

This article has reviewed learnability results that have made it possible to apply computational learning models to variable, ambiguous, and incomplete language data, arguing that probabilistic modeling has played an indispensable role in recent progress on these challenges. The most direct arguments for probabilistic models come from consistent behavioral findings demonstrating speakers' sensitivity to statistical phonological properties—such as rates of variation, gradient phonotactics, and statistically conditioned phonological alternations discussed in Section 2—together with the success of statistical models in capturing these linguistic capacities. The article has also argued, however, that successful modeling of biases shaping human generalization and learning of hidden phonological structure—such as exceptions, prosodic structure, URs and constraints—from noisy, gappy data requires quantitative models that can formalize how learners balance conflicting considerations, such as tightly fitting observed restrictions and extracting broad and natural generalizations. By modeling human learning of quantitative generalizations, the solutions to these challenges have in turn led to significant empirical discoveries about the role of phonological representations, substantive biases, and other inductive biases in shaping phonological learning and typology. As reviewed in Section 4, recent findings have shown that statistical learning of phonology is sensitive to abstract phonological representations—such as features, natural classes, and syllables—and to inductive biases, such as those favoring simpler, more phonetically natural, and more systematic patterns. The connections between psycholinguistic data, quantitative models, and typology are also starting to provide deeper insights into how learning biases may influence sound change and typology.

These exciting discoveries notwithstanding, there is still much work to be done to continue to make more realistic assumptions about the learning task, to formalize the interaction of powerful statistical learning and soft inductive biases, and to understand the relationship between learning and other factors that shape typology. Advances in hidden structure learning and in the learning of quantitative generalizations have largely proceeded independently, yet the key ingredients for integrating these approaches and modeling the learning of deeper phonological structure from natural language data are now available. This integration will undoubtedly lead to further empirical breakthroughs in our understanding of the representations and computations that underlie phonological knowledge and learning.

SUMMARY POINTS

1. Recent developments in computational phonology have made it possible to model learning from ambiguous, inconsistent, and incomplete data characteristic of natural languages.
2. Learning of quantitative generalizations in the face of noise, variability, and exceptions is an area of substantial recent progress.
3. Another area of significant recent progress is learning in the face of hidden structure and ambiguity.
4. Models that are sensitive to quantitative properties of language data, such as probabilistic models, have been indispensable to progress in both areas by providing principled ways to formalize trade-offs between conflicting pressures and to navigate ambiguity.
5. These models have made it possible to create formal links between explicit theories of learning and a rich and complex empirical base, including findings from psycholinguistics, typology, and diachrony.

6. These links have in turn led to significant empirical discoveries about the representations and computations that underlie phonological knowledge and learning.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

For valuable comments I am grateful to Dan Jurafsky, Aleksei Nazarov, Joe Pater, Kristine Yu, and members of the University of Massachusetts, Amherst, Sound Workshop.

LITERATURE CITED

Akers C. 2012. *Commitment-based learning of hidden linguistic structures*. PhD thesis, Rutgers Univ., New Brunswick, NJ

Albright A. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26:9–41

Albright A, Hayes B. 2003. Rules versus analogy in English past tenses: a computational/experimental study. *Cognition* 90:119–61

Alderete J, Tesar B. 2002. *Learning covert phonological interaction: an analysis of the problem posed by the interaction of stress and epenthesis*. RuCCS tech. rep. TR-72, Rutgers Univ., New Brunswick, NJ

Allen B, Becker M. 2015. *Learning alternations from surface forms with sublexical phonology*. Unpubl. ms., Univ. B. C., Vancouver, Can./Stony Brook Univ., Stony Brook, NY. **http://ling.auf.net/lingbuzz/002503**

Anttila A. 2007. Variation and optionality. In *The Cambridge Handbook of Phonology*, ed. P de Lacy, pp. 519–36. Cambridge, UK: Cambridge Univ. Press

Apoussidou D. 2007. *The Learnability of Metrical Phonology*. Utrecht, Neth.: LOT

Apoussidou D, Boersma P. 2003. The learnability of Latin stress. In *Proceedings from the Institute of Phonetic Sciences 25*, pp. 101–48. Amsterdam: Univ. Amsterdam

Bailey TM, Hahn U. 2001. Determinants of wordlikeness: phonotactics or lexical neighborhoods? *J. Mem. Lang.* 44:568–91

Becker M. 2009. *Phonological trends in the lexicon: the role of constraints*. PhD thesis, Univ. Mass., Amherst

Becker M, Gouskova M. 2016. Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguist. Inq.* 47:391–425

Becker M, Nevins A, Ketrez N. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87:84–125

Berent I, Steriade D, Lennertz T, Vaknin V. 2007. What we know about what we have never heard: evidence from perceptual illusions. *Cognition* 104:591–630

Berger AL, Della Pietra VJ, Della Pietra SA. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.* 22:39–71

Berko J. 1958. The child's learning of English morphology. *Word* 14:150–77

Bermúdez-Otero R. 1999. *Constraint interaction in language change: quantity in English and Germanic*. PhD thesis, Univ. Manchester, UK

Berwick RC. 1985. *The Acquisition of Syntactic Knowledge*. Artif. Intell. Ser. 16. Cambridge, MA: MIT Press

Blanchard D, Heinz J, Golinkoff R. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *J. Child Lang.* 37:487–511

Blevins J. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge, UK: Cambridge Univ. Press

Boersma P. 1997. How we learn variation, optionality, and probability. *IFA Proc.* 21:43–58

Boersma P. 2003. Review of Tesar & Smolensky (2000): *Learnability in Optimality Theory*. *Phonology* 20:436–46
Boersma P. 2011. A programme for bidirectional phonology and phonetics and their acquisition and evolution. In *Bidirectional Optimality Theory*, ed. A Benz, J Mattausch, pp. 33–72. Amsterdam: Benjamins
Boersma P, Hayes B. 2001. Empirical tests of the gradual learning algorithm. *Linguist. Inq.* 32:45–86
Boersma P, Levelt C. 2000. Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of the 30th Child Language Research Forum*, pp. 229–37. Stanford, CA: Cent. Study Lang. Inf.
Boersma P, Pater J. 2016. Convergence properties of a gradual learning algorithm for harmonic grammar. In *Harmonic Grammar and Harmonic Serialism*, ed. J McCarthy, J Pater, pp. 389–484. London: Equinox
Breteler J. 2018. *A Foot-Based Typology of Tonal Reassociation: Perspectives from Synchrony and Learnability*. Utrecht, Neth.: LOT
Calamaro S, Jarosz G. 2015. Learning general phonological rules from distributional information: a computational model. *Cogn. Sci.* 39:647–66
Chandlee J, Eyraud R, Heinz J. 2014. Learning strictly local subsequential functions. *Trans. Assoc. Comput. Linguist.* 2:491–503
Chandlee J, Heinz J. 2018. Strict locality and phonological maps. *Linguist. Inq.* 49:23–60
Chomsky N. 1981. Principles and parameters in syntactic theory. In *Explanations in Linguistics*, ed. N Hornstein, D Lightfoot, pp. 32–75. London: Longman
Clements G. 1990. The role of the sonority cycle in core syllabification. In *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, ed. J Kingston, M Beckmann, pp. 283–333. Cambridge, UK: Cambridge Univ. Press
Coetzee AW. 2009. Learning lexical indexation. *Phonology* 26:109–45
Coetzee AW, Pater J. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Nat. Lang. Linguist. Theory* 26:289–337
Coetzee AW, Pater J. 2011. The place of variation in phonological theory. In *The Handbook of Phonological Theory*, ed. J Goldsmith, J Riggle, A Yu, pp. 401–31. London: Blackwell. 2nd ed.
Coleman J, Pierrehumbert J. 1997. Stochastic phonological grammars and acceptability. arXiv:cmp-lg/9707017
Cotterell R, Peng N, Eisner J. 2015. Modeling word forms using latent underlying morphs and phonology. *Trans. Assoc. Comput. Linguist.* 3:443–47
Daland R, Hayes B, White J, Garellek M, Davis A, Norrmann I. 2011. Explaining sonority projection effects. *Phonology* 28:197–234
Daland R, Pierrehumbert JB. 2011. Learning diphone-based segmentation. *Cogn. Sci.* 35:119–55
Davidson L. 2006. Phonology, phonetics, or frequency: influences on the production of non-native sequences. *J. Phon.* 34:104–37
Della Pietra SA, Della Pietra VJ, Lafferty J. 1997. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 19:380–93
Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* 39:1–38
Dillon B, Dunbar E, Idsardi W. 2013. A single-stage approach to learning phonological categories: insights from Inuktitut. *Cogn. Sci.* 37:344–77
Dresher BE. 1999. Charting the learning path: cues to parameter setting. *Linguist. Inq.* 30:27–67
Dresher BE. 2016. Covert representations, contrast, and the acquisition of lexical accent. In *Dimensions of Phonological Stress*, ed. J Heinz, R Goedemans, H van der Hulst, pp. 231–62. Cambridge, UK: Cambridge Univ. Press
Dresher BE, Kaye JD. 1990. A computational learning model for metrical phonology. *Cognition* 34:137–95
Eisner J. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 1–8. Stroudsburg, PA: Assoc. Comput. Linguist.
Ernestus M, Baayen RH. 2003. Predicting the unpredictable: interpreting neutralized segments in Dutch. *Language* 79:5–38
Feldman NH, Griffiths TL, Morgan JL. 2009. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp. 2208–13. Austin, TX: Cogn. Sci. Soc.
Finley S, Badecker W. 2009. Artificial language learning and feature-based generalization. *J. Mem. Lang.* 61:423–37

Frisch SA, Large NR, Pisoni DB. 2000. Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *J. Mem. Lang.* 42:481–96
Gibson E, Wexler K. 1994. Triggers. *Linguist. Inq.* 25:407–54
Gildea D, Jurafsky D. 1996. Learning bias and phonological-rule induction. *Comput. Linguist.* 22:497–530
Goldwater S, Griffiths TL, Johnson M. 2009. A Bayesian framework for word segmentation: exploring the effects of context. *Cognition* 112:21–54
Goldwater S, Johnson M. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Workshop on Variation Within Optimality Theory*, pp. 113–22. Stockholm: Stockholm Univ.
Gouskova M, Becker M. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *Nat. Lang. Linguist. Theory* 31:735–65
Hayes B. 1999. Phonetically driven phonology: the role of Optimality Theory and inductive grounding. In *Formalism and Functionalism in Linguistics*, vol. 1: *General Papers*, ed. M Darnell, E Moravcsik, F Newmeyer, M Noonan, KM Wheatley, pp. 243–85. Amsterdam: Benjamins
Hayes B. 2004. Phonological acquisition in Optimality Theory: the early stages. In *Fixing Priorities: Constraints in Phonological Acquisition*, ed. R Kager, J Pater, W Zonneveld, pp. 245–91. Cambridge, UK: Cambridge Univ. Press
Hayes B. 2011. Interpreting sonority-projection experiments: the role of phonotactic modeling. In *Proceedings of the 17th International Congress of Phonetic Sciences* (*ICPhS XVII*), pp. 835–38. London: Int. Phon. Assoc.
Hayes B. 2017. Varieties of Noisy Harmonic Grammar. In *Proceedings of the 2016 Annual Meeting on Phonology*. Washington, DC: Linguist. Soc. Am. 17 pp.
Hayes B, Londe ZC. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23:59–104
Hayes B, White J. 2013. Phonological naturalness and phonotactic learning. *Linguist. Inq.* 44:45–75
Hayes B, Wilson C. 2008. A Maximum Entropy model of phonotactics and phonotactic learning. *Linguist. Inq.* 39:379–440
Hayes B, Zuraw K, Siptár P, Londe Z. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85:822–63
Heinz J. 2018. The computational nature of phonological generalizations. In *Phonological Typology*, ed. L Hyman, F Plank, pp. 1–61. Berlin: Mouton
Hughto C. 2018. Investigating the consequences of iterated learning in phonological typology. In *Proceedings of the Society for Computation in Linguistics* (*SCiL*) *2018*, pp. 182–85. Amherst, MA: Grad. Linguist. Stud. Assoc.
Inkelas S, Orgun O, Zoll C. 1997. The implications of lexical exceptions for the nature of grammar. In *Optimality Theory in Phonology: A Reader*, ed. JJ McCarthy, pp. 542–51. New York: Wiley
Itô J, Mester A. 1999. The phonological lexicon. In *The Handbook of Japanese Linguistics*, ed. N Tsujimura, pp. 62–100. Malden, MA: Blackwell
Jäger G. 2007. Maximum Entropy models and stochastic Optimality Theory. In *Architectures, Rules, and Preferences: Variation on Themes by Joan Bresnan*, ed. A Zaenen, J Simpson, T Holloway King, J Grimshaw, J Maling, C Manning, pp. 467–79. Stanford, CA: Cent. Study Lang. Inf.
Jarosz G. 2006. *Rich lexicons and restrictive grammars—maximum likelihood learning in Optimality Theory*. PhD thesis, Johns Hopkins Univ., Baltimore
Jarosz G. 2009. Restrictiveness and phonological grammar and lexicon learning. In *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society*, ed. M Elliot, J Kirby, O Sawada, E Staraki, S Yoon, pp. 125–34. Chicago: Chicago Linguist. Soc.
Jarosz G. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *J. Child Lang.* 37(spec. issue):565–606
Jarosz G. 2013. Learning with hidden structure in Optimality Theory and Harmonic Grammar: beyond robust interpretive parsing. *Phonology* 30:27–71
Jarosz G. 2015. *Expectation driven learning of phonology*. Work. pap., Univ. Mass., Amherst
Jarosz G. 2016a. Learning with violable constraints. In *The Oxford Handbook of Developmental Linguistics*, ed. J Lidz, W Snyder, J Pater. Oxford, UK: Oxford Handb. Online

Jarosz G. 2016b. Learning opaque and transparent interactions in Harmonic Serialism. In *Proceedings of the 2015 Annual Meeting on Phonology*. Washington, DC: Linguist. Soc. Am. 12 pp.

Jarosz G. 2017. Defying the stimulus: acquisition of complex onsets in Polish. *Phonology* 34:269–98

Jarosz G, Calamaro S, Zentz J. 2017. Input frequency and the acquisition of syllable structure in Polish. *Lang. Acquis.* 24:361–99

Jarosz G, Johnson JA. 2013. The richness of distributional cues to word boundaries in speech to young children. *Lang. Learn. Dev.* 9:175–210

Jarosz G, Rysling A. 2017. Sonority sequencing in Polish: the combined roles of prior bias & experience. In *Proceedings of the 2016 Annual Meeting on Phonology*. Washington, DC: Linguist. Soc. Am. 12 pp.

Jesney K, Tessier A-M. 2011. Biases in Harmonic Grammar: the road to restrictive learning. *Nat. Lang. Linguist. Theory* 29:251–90

Johnson M. 1984. A discovery procedure for certain phonological rules. In *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting of the Association for Computational Linguistics*, pp. 344–47. Stroudsburg, PA: Assoc. Comput. Linguist.

Johnson M. 2002. Optimality-Theoretic lexical functional grammar. In *The Lexical Basis of Syntactic Processing: Formal, Computational and Experimental Issues*, ed. S Stevenson, P Merlo, pp. 59–73. Amsterdam: Benjamins

Johnson M. 2008a. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the 10th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology* (*ACL SIGMORPHON*), pp. 20–27. Stroudsburg, PA: Assoc. Comput. Linguist.

Johnson M. 2008b. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp. 398–406. Stroudsburg, PA: Assoc. Comput. Linguist.

Kiparsky P. 1968. Linguistic universals and linguistic change. In *Universals in Linguistic Theory*, ed. B Emmon, RT Harms, pp. 170–202. New York: Holt, Reinhart & Winston

Kiparsky P. 1971. Historical linguistics. In *A Survey of Linguistic Science*, ed. WO Dingwall, pp. 576–642. College Park: Univ. Md. Linguist. Program

Kiparsky P. 2000. Opacity and cyclicity. *Linguist. Rev.* 17:351–66

Kirby S, Smith K, Brighton H. 2004. From UG to universals: linguistic adaptation through iterated learning. *Stud. Lang.* 28:587–607

Kisseberth LM, Kenstowicz M. 1977. *Topics in Phonological Theory*. New York: Academic

Lamont A. 2018. Decomposing phonological transformations in serial derivations. In *Proceedings of the Society for Computation in Linguistics* (*SCiL*) *2018*, pp. 91–101. Amherst, MA: Grad. Linguist. Stud. Assoc.

Le Calvez R, Peperkamp S, Dupoux E. 2007. Bottom-up learning of phonemes: a computational study. In *Proceedings of the 2nd European Cognitive Science Conference*, pp. 167–72. New York: Taylor & Francis

Legendre G, Miyata Y, Smolensky P. 1990. Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness. theoretical foundations. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pp. 388–95. Cambridge, MA: Erlbaum

Legendre G, Sorace A, Smolensky P. 2006. The Optimality Theory Harmonic Grammar connection. See Smolensky & Legendre 2006, pp. 339–402

Magri G. 2012. Convergence of error-driven ranking algorithms. *Phonology* 29:213–69

McCarthy JJ. 2000. Harmonic serialism and parallelism. In *Proceedings of the 30th Meeting of the North East Linguistics Society*, ed. M Hirotani, pp. 501–24. Amherst, MA: Grad. Linguist. Stud. Assoc.

Merchant N. 2008. *Discovering underlying forms: contrast pairs and ranking*. PhD thesis, Rutgers Univ., New Brunswick, NJ

Mohri M. 1997. Finite-state transducers in language and speech processing. *Comput. Linguist.* 23:269–311

Moore-Cantwell C, Pater J. 2016. Gradient exceptionality in Maximum Entropy grammar with lexically specific constraints. *Catalan J. Linguist.* 15:53–66

Moore-Cantwell C, Staubs RD. 2014. Modeling morphological subgeneralizations. In *Proceedings of the 2013 Annual Meeting on Phonology*. Washington, DC: Linguist. Soc. Am. 11 pp.

Moreton E. 2008. Analytic bias and phonological typology. *Phonology* 25:83–127

Moreton E, Pater J. 2012a. Structure and substance in artificial-phonology learning. Part II: Substance. *Lang. Linguist. Compass* 6:686–701

Moreton E, Pater J. 2012b. Structure and substance in artificial-phonology learning. Part I: Structure. *Lang. Linguist. Compass* 6:702–18
Moreton E, Pater J, Pertsova K. 2015. Phonological concept learning. *Cogn. Sci.* 2015:1–66
Nazarov A, Jarosz G. 2017. Learning parametric stress without domain-specific mechanisms. In *Proceedings of the 2016 Annual Meeting on Phonology*. Washington, DC: Linguist. Soc. Am. 12 pp.
Nazarov AI. 2016. *Extending hidden structure learning: features, opacity, and exceptions*. PhD thesis, Univ. Mass., Amherst
Nazarov AI. 2018. Learning within- and between-word variation in probabilistic OT grammars. In *Proceedings of the 2017 Annual Meeting on Phonology*. Washington, DC: Linguist. Soc. Am. 12 pp.
Nazarov AI, Pater J. 2017. Learning opacity in stratal Maximum Entropy grammar. *Phonology* 34:299–324
Ohala JJ. 1993. The phonetics of sound change. In *Historical Linguistics: Problems and Perspectives*, ed. C Jones, pp. 237–78. London: Longman
O'Hara C. 2017. How abstract is more abstract? Learning abstract underlying representations. *Phonology* 34:325–45
O'Hara C. 2018. *Emergent learning bias and the underattestation of simple patterns*. Work. pap., Univ. South. Calif., Pasadena
Oncina J, García P, Vidal E. 1993. Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* 15:448–58
Pater J. 2008. Gradual learning and convergence. *Linguist. Inq.* 39:334–45
Pater J. 2010. Morpheme-specific phonology: constraint indexation and inconsistency resolution. In *Phonological Argumentation: Essays on Evidence and Motivation*, ed. S Parker, pp. 123–54. London: Equinox
Pater J. 2012. Emergent systemic simplicity (and complexity). *McGill Univ. Work. Pap. Linguist.* 22:1
Pater J, Jesney K, Staubs RD, Smith B. 2012. Learning probabilities over underlying representations. In *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology* (*SIGMORPHON2012*), pp. 62–71. Stroudsburg, PA: Assoc. Comput. Linguist.
Pearl LS. 2011. When unbiased probabilistic learning is not enough: acquiring a parametric system of metrical phonology. *Lang. Acquis.* 18:87–120
Peperkamp S, Le Calvez R, Nadal JP, Dupoux E. 2006. The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition* 101:B31–41
Pierrehumbert JB. 2001. Exemplar dynamics: word frequency, lenition and contrast. In *Frequency and the Emergence of Linguistic Structure*, ed. JL Bybee, PJ Hopper, pp. 137–57. Amsterdam: Benjamins
Prickett B. 2018a. *Experimental evidence for biases in phonological rule interaction*. Poster presented at the 16th Conference on Laboratory Phonology, Lisbon
Prickett B. 2018b. Similarity-based phonological generalization. In *Proceedings of the Society for Computation in Linguistics* (*SCiL*) *2018*, pp. 193–96. Amherst, MA: Grad. Linguist. Stud. Assoc.
Prickett B. 2018c. Complexity and naturalness biases in phonotactics: Hayes and White (2013) revisited. In *Proceedings of the 2017 Annual Meeting on Phonology*. Washington, DC: Linguist. Soc. Am. 6 pp.
Prince A. 2010. *Counting parses*. Work. pap., Rutgers Univ., New Brunswick, NJ
Prince A, Smolensky P. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. New York: Wiley
Prince A, Tesar B. 2004. Learning phonotactic distributions. In *Fixing Priorities: Constraints in Phonological Acquisition*, ed. R Kager, J Pater, W Zonneveld, pp. 245–91. Cambridge, UK: Cambridge Univ. Press
Rasin E, Berger I, Katzir R. 2015. *Learning rule-based morpho-phonology*. Work. pap., MIT, Cambridge, MA
Rasin E, Katzir R. 2016. On evaluation metrics in Optimality Theory. *Linguist. Inq.* 47:235–82
Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65:386–408
Selkirk E. 1982. The syllable. In *The Structure of Phonological Representations*, ed. H van der Hulst, N Smith, pp. 337–84. Dordrecht, Neth.: Foris
Shih SS. 2018. Learning lexical classes from variable phonology. In *Proceedings of the 2nd Asian Junior Linguists Conference*. Tokyo: Int. Christ. Univ. 15 pp.
Smith BW. 2015. *Phonologically conditioned allomorphy and UR constraints*. PhD thesis, Univ. Mass., Amherst
Smith BW, Pater J. 2017. *French schwa and gradient cumulativity*. Work. pap., Univ. Calif., Berkeley/Univ. Mass., Amherst

Smolensky P, Goldrick M. 2016. *Gradient symbolic representations in grammar: the case of French liaison*. Work, pap., Johns Hopkins Univ., Baltimore/Northwest. Univ., Evanston, IL
Smolensky P, Legendre G, ed. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, MA: MIT Press
Soderstrom M, Mathis D, Smolensky P. 2006. Abstract genomic encoding of Universal Grammar in Optimality Theory. See Smolensky & Legendre 2006, pp. 403–71
Solomonoff RJ. 1964. A formal theory of inductive inference. Parts I and II. *Inf. Control* 7:1–22, 224–54
Stanton J. 2016. Learnability shapes typology: the case of the midpoint pathology. *Language* 92:753–91
Staubs RD. 2014. *Computational modeling of learning biases in stress typology*. PhD thesis, Univ. Mass., Amherst
Staubs RD, Pater J. 2016. Learning serial constraint-based grammars. In *Harmonic Grammar and Harmonic Serialism*, ed. JJ McCarthy, J Pater, pp. 369–88. London: Equinox
Steriade D. 2008. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. In *The Nature of the Word: Studies in Honor of Paul Kiparsky*, ed. K Hanson, S Inkelas, chapter 7. Cambridge, MA: MIT Press
Tesar BB. 1995. *Computational optimality theory*. PhD thesis, Univ. Colo., Boulder
Tesar BB. 2006. Faithful contrastive features in learning. *Cogn. Sci.* 30:863–903
Tesar BB. 2013. *Output-Driven Phonology: Theory and Learning*. Cambridge, UK: Cambridge Univ. Press
Tesar BB, Prince A. 2007. Using phonotactics to learn phonological alternations. In *Proceedings of the 39th Conference of the Chicago Linguistics Society*, pp. 209–37. Chicago: Chicago Linguist. Soc.
Tesar BB, Smolensky P. 1998. Learnability in Optimality Theory. *Linguist. Inq.* 29:229–68
Tesar BB, Smolensky P. 2000. *Learnability in Optimality Theory*. Cambridge, MA: MIT Press
Tessier A-M. 2009. Frequency of violation and constraint-based phonological learning. *Lingua* 119:6–38
Vitevitch MS, Luce PA. 2004. A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behav. Res. Methods Instrum. Comput.* 36:481–87
Wexler K, Culicover P. 1980. *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press
White J. 2017. Accounting for the learnability of saltation in phonological theory: a maximum entropy model with a P-map bias. *Language* 93:1–36
Wilson C. 2006. Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cogn. Sci.* 30:945–82
Wilson C, Gallagher G. 2018. Accidental gaps and surface-based phonotactic learning: a case study of South Bolivian Quechua. *Linguist. Inq.* 49:610–23
Yang CD. 2002. *Knowledge and Learning in Natural Language*. Oxford, UK: Oxford Univ. Press
Yu KM. 2017. Advantages of constituency: computational perspectives on Samoan word prosody. In *Proceedings of the 22nd International Conference on Formal Grammar*, pp. 105–24. Berlin: Springer
Zuraw K. 2000. *Patterned exceptions in phonology*. PhD thesis, Univ. Calif., Los Angel.

Annual Review
of Linguistics

Volume 5, 2019

Contents

Errata

An online log of corrections to *Annual Review of Linguistics* articles may be found at http://www.annualreviews.org/errata/linguistics