Computational Modeling of Phonological Learning

Gaja Jarosz

University of Massachusetts Amherst

jarosz@linguist.umass.edu

Running Title: Modeling Phonological Learning

Department of Linguistics
Integrative Learning Center N410
University of Massachusetts
Amherst, MA 01003-1100

April 30, 2018

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 the 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 threads 1) the development of models capable of learning the quantitative, noisy, and inconsistent patterns that are characteristic of naturalistic data, and 2) 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 discoveries they have led to.

Modeling Phonological Learning

1. INTRODUCTION

Recent advances in computational modeling of phonological learning have had a transformative impact on the field. 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. Computational models have led to significant discoveries about the fundamental characteristics of the human language acquisition device: how language knowledge is represented and acquired and what limits 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 computational models of learning provide. This link makes it possible to test theoretical assumptions by comparing the predictions of computational models to measurable linguistic behavior in psycholinguistic experiments, typological evidence, and empirical observations about language change and loanword adaptation.

Computational modeling of phonological learning has become an essential tool of modern phonological research. It complements the rise of experimental work 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 has been made possible by modeling developments that can be broadly classified into two strands 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 language 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 more 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, it is only through detailed comparison of the quantitative patterns in natural language data and the generalizations learners infer on the basis of that data that 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 group of advances in computational modeling involves the learning of hidden linguistic structure, which is an intrinsic property of language at various levels of representation. Hidden structure includes all representations that learners must infer but which cannot be directly observed in the learning data. Depending on theoretical assumptions, hidden structure in phonology may include metrical feet, underlying representations, 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.

Lastly, Section 4 reviews significant 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 the next focuses on incompleteness, but these are two sides of the same coin: ambiguity. Ambiguity means that there are multiple interpretations, 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 greatest 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 on many forms. The child learning their first language is not told which data tokens are errors that should be ignored, nor 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 (see Anttila 2007; Coetzee & Pater 2011 for reviews). 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 sort 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). 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 where phonologically similar words or morphemes behave inconsistently in the same phonological environments.

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

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

requires sensitivity to quantitative properties of the data, and therefore much of the 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 commonly utilized 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 underlying representation. 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 underlying representation. 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 non-negative 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, post-consonantal [t] in three environments (pre-pausal (i), pre-consonantal (ii), and pre-vocalic (iii)). There is one constraint that penalizes post-consonantal [t] (*CT), one general MAX constraint, and two contextual variants of MAX, one specific to the pre-vocalic 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 the first tableau (i), MAX-FIN (2) and MAX (3) together assign a harmony of -5 to the deletion candidate (/Ct/, [C_]), while the faithful candidate (/Ct/, [Ct]) violates only *CT, receiving a harmony of -4. Thus, in the pre-pausal context, the faithful candidate has higher harmony and is preferred according to these weights. In the second competition representing the preconsonantal context (ii), the deletion candidate (/CtC/, [C_C]) violates only the general MAX (3), making it more harmonic than the faithful candidate (/CtC/, [CtC]), while in the final tableau representing the pre-vocalic context (iii), 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 (i), the faithful candidate has probability $\frac{\exp(-4)}{\exp(-4)+\exp(-5)} \cong 73.1\%$ while the deletion candidate gets $\frac{\exp(-5)}{\exp(-4)+\exp(-5)} \cong 26.9\%$. With these weights, the probabilities of the faithful candidates in the second (ii) and third tableau (iii) are roughly 26.9% and 50%.

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 modelling free variation in the literature using these frameworks (see e.g. Boersma & Hayes 2001; Coetzee & Pater 2008; Goldwater & Johnson 2003).

2.2 Learning Free Variation

Numerous successful algorithms have been developed for learning categorical OT rankings and HG weightings from full structural descriptions[2] (Boersma & Pater 2016; Goldwater & Johnson 2003; Jäger 2007; Magri 2012; Soderstrom et al. 2006; Tesar 1995). There are a number of online and batch algorithms for both OT and HG that are guaranteed to find a categorical target grammar for any set of input-output pairs, as long as such a target grammar exists. The online, error-driven constraint demotion (EDCD) algorithm (Tesar 1995) forms the basis for a number of frequency-sensitive models. 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.

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 input-output pair (x, y), and the learner incorrectly selects (x, y') as the winning candidate, an error. The learner

[2] Learning from full structural descriptions means the learner is provided with access to all representations referenced by constraints, including hidden representations. Moving beyond this simplifying assumption is the focus of the next section.

compares the constraint violations of the observed (x, y) to the violations of the error (x, y'), slightly demoting constraints that favor the error and slightly promoting constraints that favor the observed form. 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. Each time the learner observes (/kɑst/, [kɑst]), it must predict [kɑst] as the output, and [kɑs] will be treated as an error, while 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. In this way, systematic free variation 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 outside of linguistics in a variety of machine learning and natural language processing contexts, and 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 for these models, and they are guaranteed to find the weights that best fit the observed distribution. Jäger (2007) shows that the stochastic gradient descent 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 probabilities 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.

2.3 Gradient Phonotactics

The MaxEnt, Stochastic OT, and Noisy HG frameworks can all be used for modeling graded acceptability and phonotactics as well (Boersma & Hayes 2001; Coetzee & Pater 2008; Hayes & Wilson 2008)[3]. The most common approach follows Hayes and 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. The predicted probabilities of various word forms can then be numerically transformed and correlated with acceptability scales or other behavioral measures. The Hayes and Wilson Phonotactic Learner (Hayes & Wilson 2008) has been especially broadly applied in recent years and has performed well in predicting experimentally elicited phonotactic scales (see e.g. Albright 2009; Daland et al. 2011). These applications will be discussed further in Section 4. In addition to dealing with inconsistency, the Phonotactic Learner also takes on a hidden structure learning problem, learning constraints, which will be discussed further in Section 3.

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

2.4 Classes, Exceptions, and Lexicalized Variation

Learning classes, exceptions, and lexicalized variation faces 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 just 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 has 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 work on learning lexical exceptionality in a constraint-based framework (Becker 2009; Coetzee 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.

While the RCD-based exceptionality approach can deal with one kind of inconsistency (exceptions), it cannot cope with learning data that has exceptions and other kinds of inconsistency or ambiguity, like variability or hidden structure. A variety of approaches to learning exceptions or classes in the face of variability have recently been developed by extending frequency-sensitive approaches (Nazarov 2016, 2018; Pater et al. 2012; Shih 2018). While the details vary, all 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 work focuses on the gradient productivity of morpho-phonological transformations using rules (Albright & Hayes 2003) and constraints (Allen & Becker 2015; Becker & Gouskova 2016; Moore-Cantwell & Staubs 2014).

An empirical and theoretical problem of particular interest in recent modeling work is that of gradient, or patterned, exceptionality. A number of experimental studies across multiple languages have now shown that modeling speaker's 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 models discussed earlier 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 these models generally learn language-wide patterns more quickly than lexically-specific patterns (Moore-Cantwell & Pater 2016; Pater et al. 2012; Zuraw 2000). This is because of these models' sensitivity to frequency: all the learning data support language-wide patterns, while support for lexically-specific patterns occurs rarely, only when the particular lexical item is observed. This allows the models to learn language-wide statistical preferences early on, before the memorized properties of individual lexical items begin to dominate production and cause learning to slow.

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 success. 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. What 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 will 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 discuss the novels insights on long-standing debates that recent modeling work 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 (Prince 2010), it grows exponentially or worse with the number of words, features, or constraints. In the case of learning abstract underlying representations, 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 and Hayes 2003). The learner must therefore 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 learning phonological mappings and underlying representations, 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 must be blamed. 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 necessary update to the grammar. As discussed in Section 2.1, this learning sub-problem has been solved. 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 get anywhere with hidden structure learning.

Another source of ambiguity in hidden structure learning is the relative breadth or narrowness of inferred generalizations. The subset problem discussed earlier for exceptionality arises when learning underlying representations or any other lexically-specific properties[4]. A related issues 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 to appear)). 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

[4] 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).

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, substantive and quantitative factors influence this process – is an important area of ongoing research.

3.2 Approaches & 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. Numerous models discussed in this section build on well-studied techniques in machine learning like 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 like variability and exceptionality. Frequency-sensitive learning approaches can also provide a way to 'break into' the chicken and egg ambiguity that hidden structure creates. Modeling work 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 & 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 like 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 to appear; Wilson & Gallagher to appear).

3.3 Significant Results

3.3.1 Prosodic Structure

One of the most well-studied hidden structure learning problems in phonology is that of metrical structure. While metrical structure has been given particular emphasis, 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 learning models can address fundamental questions about innate linguistic knowledge. To tackle the overwhelming ambiguity created by metrical footing, pioneering work (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. In addition to innate cues, Dresher and Kaye furthermore hypothesized that successful learning requires that parameters have default settings and an inherent ordering. Pearl (2011) recently 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 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 Optimality Theory. Tesar & Smolensky (1998, 2000) proposed a parsing strategy called Robust Interpretive Parsing (RIP) that allowed 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 earlier: they use their current grammar to guess at 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 to the learner's own production, which is the optimal candidate among all possible stress assignments for /tɛlɛfɔn/. If there is a mismatch, the constraint ranking is updated as usual based on the constraint violations of both candidates.

Building on Tesar & Smolensky's proposal, the parsing approach 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 the approach to OT-GLA, while Boersma & Pater (2016) extended it to 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. In subsequent work, 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. She showed these strategies substantially out-performed RIP and leveled the performance of the OT and HG learning models, revealing that the OT disadvantage discovered by Boersma & Pater was due to properties specific to RIP rather than OT per se. In follow-up work, 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, like lexical representations and derivations, discussed next.

3.3.2 Lexical Representations

Much of the earliest work on learning underlying representations focused on lexical accent. Even when learning is restricted to learning underlying features of observed segments – that is, if insertion and deletion mappings are not considered – the space of possible underlying representations is exponentially large. To be computationally feasible, models must therefore find efficient ways to navigate the exponential space (Jarosz 2015; Merchant 2008; Tesar 2013) or restrict the feature values or forms considered by the learner to those observed on the surface (Hayes 2004; Pater et al. 2012; Tesar 2006). A variety of representational approaches have been developed for modeling lexical properties. Some assume the traditional underlying representation 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) can also be applied to the learning of lexical representations[5].

The computational pressures are 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 to appear). As discussed earlier, learning of deletion mappings opens the door to a potentially infinite space of abstract underlying representations. To model this aspect of phonological learning, assumptions about the range of lexical options available to the learner must be made explicit. Work 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 work is beginning to provide new arguments for both abstract (O'Hara 2017) and concrete lexical representations (Allen & Becker 2015). There is potential to make the trade-offs between more abstract lexical representations and the ability to (efficiently) learn attested kinds of alternations explicit by applying, extending, and testing the current range of learning models.

3.3.3 Derivations & Intermediate Representations

Learning of serial derivations is probably the least well understood learning problem in phonology, and 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 work 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

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

learning of underlying representations and ordered rules from paradigmatic information, but this procedure made strong simplifying assumptions about types of rules and interactions allowed – for example, insertion and deletion were 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, see Mohri (1997)) and is guaranteed to learn the target mapping in the limit (Oncina et al. 1993). Gildea & Jurafsky showed that the problem arises due to lack of sufficient restrictions on generalization. As discussed earlier, naturalistic data does not provide every combination of features or segments that instantiate a rule or pattern. Without biases favoring more natural[6] phonological rules, the algorithm fails to generalize appropriately to unseen data. More recent work on subregular formalizations of phonology have investigated even tighter restrictions on permissible mappings (Chandlee et al. 2014; Chandlee & Heinz 2018). However, these learning procedures still assume the learner observes input-output pairs and all combinations of segments that instantiate the pattern.

There are also recent frequency-sensitive approaches. Rasin, Berger, & Katzir (2015) pursue an approach using principles of minimum description length (Solomonoff 1964) to learn both underlying representations 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

[6] Gildea & Jurafsky proposed three biases that improved learning outcomes: faithfulness, community, and context.

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).

3.3.4 Constraints

In 2008, Hayes & Wilson 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. Prior to this work, 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 under-attested 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 work, discussed in the next section, investigating computationally and experimentally what biases are required to account for human learning and generalization.

To formalize under-attestation 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 to the expected number of violations of that constraint given the current grammar and weights. The observed violations are straightforward to calculate: this 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. It is only by noticing that a constraint like *#ŋ correctly rules out unattested forms that would otherwise be predicted that the learner can induce this constraint and weight it highly. Hayes and Wilson use finite-state methods to estimate the expected violations efficiently.

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 to appear). As discussed earlier, 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.

4. Modeling Human Learning, Generalization, and Typology

This article has argued that sensitivity to quantitative patterns in natural language data is essential for modeling variation and exceptionality and for tackling learning challenges posed by hidden structure. This final section reviews some of the most significant discoveries about human learning and generalization that such models have revealed. Natural language has statistical information, and learners are sensitive to this information – it is only by modeling learners' sensitivity to this information that we can 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. At the same time, the integration of modeling and behavioral work 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 datasets of phonetically or phonemically transcribed words or paradigms in the target language. By comparing predictions of models making different theoretical or representational assumptions, it is possible to make inferences about the likely contents of the human language acquisition device. Wug tests (Berko 1958) are used to study speakers' productive knowledge of morpho-phonological alternations in their language and can be compared to models that generate predictions about alternations. Another way to probe speakers' knowledge is by comparing acceptability judgments on phonotactic patterns or alternations to 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 work using quantitative models demonstrated the success of stochastic grammars capable of extracting abstract generalizations from the lexicon (Albright & Hayes 2003; Boersma & Hayes 2001; Coleman & Pierrehumbert 1997). For example, Albright and Hayes (2003) showed that abstract rules better capture learning of morpho-phonological alternations than analogical models that directly compute overall similarity with the lexicon. A subsequent series of studies demonstrated that sensitivity to abstract phonological representations like 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 could successfully predict speakers' gradient preferences for higher sonority rises (Clements 1990; Selkirk 1982).

Perhaps the most broadly investigated question in recent modeling work concerns the role of phonetic naturalness and substantive bias (Becker et al. 2011; Berent et al. 2007; Davidson 2006; Hayes & Londe 2006; Hayes et al. 2009; Hayes & White 2013; Jarosz & Rysling 2017; O'Hara 2018; Prickett 2018a,b). 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 UG (e.g. Prince & Smolensky 2004), 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)? While more work is still needed, the emerging view that recent modeling work supports is that phonetic substance 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 to some, it is in conflict 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 2001), making it harder to learn high weights for phonetically unmotivated constraints (White 2017; Wilson 2006). However, there is still much work 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 cross-linguistically.

To summarize, modeling of first language acquisition has provided evidence for the role of abstract phonological representations and soft substantive biases.[7]

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 in natural languages. In the ALL context, evidence for substantive bias (Finley & Badecker 2009; White 2017; Wilson 2006) has been rather weak and mixed (Moreton & Pater 2012a). Since evidence from first language acquisition has demonstrated sensitivity to phonetic naturalness, this discrepancy is likely due to the differences between first language acquisition and artificial language learning. 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.

[7] 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).

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 and Pater show that it is important to keep the effect of complexity in mind when examining other pressures since complexity and naturalness are often correlated (see also Prickett 2018b). Formalizing simplicity and comparing its effects in linguistic and non-linguistic domains has also been investigated in recent modeling work (Moreton et al. 2015).

4.3 Modeling Diachrony & Typology

There is also a growing body of work using quantitative models of phonological learning to investigate soft learning biases that could be responsible for cross-linguistic tendencies and universals[8]. 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 universal grammar (UG). This perspective precludes the possibility of modeling cross-linguistic tendencies rather than strict universals and overlooks pressures besides UG that may be involved in shaping the observed typology, such as domain-general learning biases. As discussed earlier, models of phonological learning make predictions about the relative ease of learning of various patterns: some patterns are learned more quickly or require less data than others, and when there is hidden structure,

[8] Work 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 I have argued are essential to modeling human learning in the face of noise and ambiguity (though see Lamont 2018; Yu 2017).

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 were already 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 cross-linguistic tendencies. These results have broad implications for linguistic theory since they show that biases inherent to statistical learning can systematically skew the typological predictions that follow from theoretical assumptions.

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.

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.

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. The advances on hidden structure learning and on 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 no doubt 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 one 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, like probabilistic models, have been indispensable to the 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 have led to significant empirical discoveries about the representations and computations that underlie phonological knowledge and learning.

ACKNOWLEDGMENTS

For valuable discussion I am grateful to 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*

Albright A. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology*. 26(01):9–41

Albright A, Hayes B. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*. 90(2):119–61

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

Allen B, Becker M. 2015. Learning alternations from surface forms with sublexical phonology. *Unpubl. Manuscr. Univ. Br. Columbia Stony Brook Univ. Available Lingbuzz002503*

Anttila A. 2007. Variation and Optionality. In *The Cambridge Handbook of Phonology*, ed. P de Lacy, pp. 519–36. Cambridge University Press

Apoussidou D. 2007. *The learnability of metrical phonology*

Apoussidou D, Boersma P. 2003. The learnability of Latin Stress. *IFA Proc.* 25:101–48

Bailey TM, Hahn U. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *J. Mem. Lang.* 44(4):568–91

Becker M. 2009. *Phonological Trends in the Lexicon: The Role of Constraints*. University of Massachusetts, Amherst.

Becker M, Gouskova M. 2016. Source-Oriented Generalizations as Grammar Inference in Russian Vowel Deletion. *Linguist. Inq.* 47(3):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(1):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(3):591–630

Berger AL, Della Pietra VJ, Della Pietra SA. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.* 22(1):39–71

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

Bermúdez-Otero R. 1999. *Constraint interaction in language change: quantity in English and Germanic*. PhD Thesis thesis. University of Manchester

Berwick RC. 1985. *The Acquisition of Syntactic Knowledge*, Vol. 16. MIT press

Blanchard D, Heinz J, Golinkoff R. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *J. Child Lang.* First View:1–25

Blevins J. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University 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(3):436–46

Boersma P. 2011. A programme for bidirectional phonology and phonetics and their acquisition and evolution. *Bidirectional Optim. Theory*. 180:33

Boersma P, Hayes B. 2001. Empirical Tests of the Gradual Learning Algorithm. *Linguist. Inq.* 32(1):45–86

Boersma P, Levelt C. 2000. Gradual Constraint-Ranking Learning Algorithm Predicts Acquisition Order. In *Proceedings of 30th Child Language Research Forum*, pp. 229–37. Stanford, California: CSLI

Boersma P, Pater J. 2016. Convergence Properties of a Gradual Learning Algorithm for Harmonic Grammar. In *Harmonic Grammar and Harmonic Serialism*, eds. J McCarthy, J Pater. London: Equinox Press

Breteler J. 2018. *A Foot-Based Typology of Tonal Reassociation: Perspectives from Synchrony and Learnability*. The Netherlands: LOT

Calamaro S, Jarosz G. 2015. Learning general phonological rules from distributional information: A computational model. *Cogn. Sci.* 39(3):647–666

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(1):23–60

Chomsky N. 1981. Principles and parameters in syntactic theory. *Explan. Linguist.* 32–75

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.*, eds. J Kingston, M Beckmann, pp. 283–333. Cambridge: Cambridge University Press

Coetzee A, Pater J. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Nat. Lang. Linguist. Theory*. 26(2):289–337

Coetzee A, Pater J. 2011. The place of variation in phonological theory. In *The Handbook of Phonological Theory*, eds. J Goldsmith, J Riggle, A Yu, pp. 401–31. Blackwell. 2nd ed.

Coetzee AW. 2009. Learning lexical indexation. *Phonology*. 26(1):109–45

Coleman J, Pierrehumbert J. 1997. Stochastic phonological grammars and acceptability. *ArXiv Prepr. Cmp-Lg9707017*

Cotterell R, Peng N, Eisner J. 2015. Modeling word forms using latent underlying morphs and phonology. *Trans. Assoc. Comput. Linguist.* 3(1):

Daland R, Hayes B, White J, Garellek M, Davis A, Norrmann I. 2011. Explaining sonority projection effects. *Phonology*. 28(02):197–234

Daland R, Pierrehumbert JB. 2011. Learning Diphone-Based Segmentation. *Cogn. Sci.* 35(1):119–55

Davidson L. 2006. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *J. Phon.* 34(1):104–37

Della Pietra SA, Della Pietra VJ, Lafferty J. 1997. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 19(4):380–93

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B Methodol.* 39(1):1–38

Dillon B, Dunbar E, Idsardi W. 2013. A Single-Stage Approach to Learning Phonological Categories: Insights From Inuktitut. *Cogn. Sci.* 37(2):344–77

Dresher BE. 1999. Charting the Learning Path: Cues to Parameter Setting. *Linguist. Inq.* 30(1):27–67

Dresher BE. 2016. Covert representations, contrast, and the acquisition of lexical accent. *Dimens. Phonol. Stress*. 231

Dresher BE, Kaye JD. 1990. A computational learning model for metrical phonology. *Cognition*. 34(2):137–95

Ernestus M, Baayen RH. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*. 5–38

Feldman NH, Griffiths TL, Morgan JL. 2009. Learning phonetic categories by learning a lexicon. *Proc. 31st Annu. Conf. Cogn. Sci. Soc.*, pp. 2208–13. Austin, TX: Cognitive Science Society

Finley S, Badecker W. 2009. Artificial language learning and feature-based generalization. *J. Mem. Lang.* 61(3):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(4):481–96

Gibson E, Wexler K. 1994. Triggers. *Linguist. Inq.* 25(3):407–54

Gildea D, Jurafsky D. 1996. Learning bias and phonological-rule induction. *Comput. Linguist.* 22(4):497–530

Goldwater S, Griffiths TL, Johnson M. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*. 112(1):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*. Stockholm University

Gouskova M, Becker M. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *Nat. Lang. Linguist. Theory*. 31(3):735–765

Hayes B. 1999. Phonetically driven phonology: The role of Optimality Theory and inductive grounding. In *Formalism and Functionalism in Linguistics, Volume 1: General Papers*, eds. M Darnell, E Moravcsik, F Newmeyer, M Noonan, KM Wheatley, pp. 243–85. Amsterdam: John Benjamins

Hayes B. 2011. Interpreting sonority-projection experiments: the role of phonotactic modeling. *Proc. 17th Int. Congr. Phon. Sci.*, pp. 835–38

Hayes B. 2017. Varieties of Noisy Harmonic Grammar. *Proc. Annu. Meet. Phonol.* 4(0):

Hayes B, Londe ZC. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology*. 23(01):59–104

Hayes B, White J. 2013. Phonological naturalness and phonotactic learning. *Linguist. Inq.* 44(1):45–75

Hayes B, Wilson C. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguist. Inq.* 39(3):379–440

Hayes B, Zuraw K, Siptár P, Londe Z. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language*. 85(4):822–63

Hayes BC. 2004. Phonological Acquisition in Optimality Theory: the Early Stages. In *Fixing Priorities: Constraints in Phonological Acquisition*, eds. R Kager, J Pater, W Zonneveld, pp. 245–91. Cambridge: Cambridge University Press

Heinz J. 2018. The computational nature of phonological generalizations. In *Phonological Typology*, eds. L Hyman, F Plank. Mouton

Hughto C. 2018. Investigating the Consequences of Iterated Learning in Phonological Typology. *Proc. Soc. Comput. Linguist.* 1:182–85

Inkelas S, Orgun O, Zoll C. 1997. The implications of lexical exceptions for the nature of grammar. *Optim. Theory Phonol. Read.* 542–551

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*, eds. A Zaenen, J

Simpson, T Holloway King, J Grimshaw, J Maling, C Manning, pp. 467–79. Stanford: CSLI Publications

Jarosz G. 2006. *Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory*

Jarosz G. 2009. Restrictiveness and Phonological Grammar and Lexicon Learning. In *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society*, Vol. 43, eds. M Elliot, J Kirby, O Sawada, E Staraki, S Yoon, pp. 125–34. Chicago Linguistics Society

Jarosz G. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *J. Child Lang. Spec. Issue Comput. Models Child Lang. Learn.* 37(3):565–606

Jarosz G. 2013. Learning with Hidden Structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology*. 30(1):27–71

Jarosz G. 2015. *Expectation Driven Learning of Phonology*. Work. Pap., University of Massachusetts, Amherst

Jarosz G. 2016a. Computational Models of Learning with Violable Constraints. In *The Oxford Handbook of Developmental Linguistics*, eds. J Lidz, W Snyder, J Pater. Oxford University Press

Jarosz G. 2016b. Learning opaque and transparent interactions in Harmonic Serialism. *Proc. Annu. Meet. Phonol.* 3:

Jarosz G. 2017. Defying the stimulus: acquisition of complex onsets in Polish. *Phonology*. 34(2):269–98

Jarosz G, Calamaro S, Zentz J. 2017. Input frequency and the acquisition of syllable structure in Polish. *Lang. Acquis.* 24(4):361–99

Jarosz G, Johnson JA. 2013. The Richness of Distributional Cues to Word Boundaries in Speech to Young Children. *Lang. Learn. Dev.* 9(2):175–210

Jarosz G, Rysling A. 2017. Sonority Sequencing in Polish: the Combined Roles of Prior Bias & Experience. *Proc. Annu. Meet. Phonol.* 4(0):

Jesney K, Tessier A-M-. M. 2011. Biases in Harmonic Grammar: the road to restrictive learning. *Nat. Lang. Linguist. Theory*. 29(1):251–90

Johnson M. 1984. A discovery procedure for certain phonological rules. *Proc. Tenth Int. Conf. Comput. Linguist.*, pp. 344–47

Johnson M. 2002. Optimality-Theoretic Lexical Functional Grammar. In *The Lexical Basis of Suntactic Processing: Formal, Computational and Experimental Issues*, eds. S Stevenson, P Merlo, pp. 59–73. Amsterdam: John Benjamins

Johnson M. 2008a. Unsupervised Word Segmentation for Sesotho Using Adaptor Grammars. *Proc. 10th Meet. ACL SIGMORPHON*, pp. 20–27. Columbus, OH: Association of Computational Linguistics

Johnson M. 2008b. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. *Proc. 46th Annu. Meet. Assoc. Comput. Linguist.*, pp. 398–406. Association for Computational Linguistics

Kiparsky P. 1968. Linguistic universals and linguistic change. In *Universals in Linguistic Theory*, eds. Bach, 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: University of Maryland Linguistics Program

Kiparsky P. 2000. Opacity and cyclicity. *Linguist. Rev.* 17(2–4):351–366

Kirby S, Smith K, Brighton H. 2004. From UG to universals: Linguistic adaptation through iterated learning. *Stud. Lang.* 28(3):587–607

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

Lamont A. 2018. Decomposing phonological transformations in serial derivations. *Proc. Soc. Comput. Linguist.* 1(1):91–101

Le Calvez R, Peperkamp S, Dupoux E. 2007. Bottom-up learning of phonemes: A computational study. *Proc. Second Eur. Cogn. Sci. Conf.*, pp. 167–72

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 Twelfth Annual Conference of the Cognitive Science Society*, pp. 388–95. Cambridge, MA: Lawrence Erlbaum

Legendre G, Sorace A, Smolensky P. 2006. The Optimality Theory Harmonic Grammar Connection. In *The Harmonic Mind : From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, Mass.: MIT Press

Magri G. 2012. Convergence of error-driven ranking algorithms. *Phonology*. 29(02):213–69

McCarthy JJ. 2000. Harmonic serialism and parallelism. In *Proceedings of the North East Linguistics Society 30*, ed. M Hirotani, pp. 501–24. Amherst, MA: GLSA Publications

Merchant N. 2008. *Discovering Underlying Forms: Contrast Pairs and Ranking*. PhD Thesis thesis. Rutgers University, New Brunswick, NJ

Mohri M. 1997. Finite-state transducers in language and speech processing. *Comput. Linguist.* 23(2):269–311

Moore-Cantwell C, Pater J. 2016. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. *Catalan J. Linguist.* 15(0):53–66

Moore-Cantwell C, Staubs RD. 2014. Modeling Morphological Subgeneralizations. *Proc. Annu. Meet. Phonol.* 1(1):

Moreton E. 2008. Analytic bias and phonological typology. *Phonology*. 25(01):83–127

Moreton E, Pater J. 2012a. Structure and substance in artificial-phonology learning, Part II: Substance. *Lang. Linguist. Compass*. 6(11):686–701

Moreton E, Pater J. 2012b. Structure and substance in artificial-phonology learning, Part I: Structure. *Lang. Linguist. Compass*. 6(11):702–218

Moreton E, Pater J, Pertsova K. 2015. Phonological Concept Learning. *Cogn. Sci.* 1–66

Nazarov A. 2016. Extending Hidden Structure Learning: Features, Opacity, and Exceptions. *Dr. Diss.*

Nazarov A. 2018. Learning within- and between-word variation in probabilistic OT grammars. In *Proceedings of the Annual Meeting on Phonology 2017*

Nazarov A, Pater J. 2017. Learning opacity in Stratal Maximum Entropy Grammar. *Phonology*. 34(2):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(2):325–345

O'Hara C. 2018. *Emergent learning bias and the underattestation of simple patterns*. Work. Pap., USC

Oncina J, García P, Vidal E. 1993. Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* 15(5):448–458

Pater J. 2008. Gradual Learning and Convergence. *Linguist. Inq.* 39(2):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). In *Proceedings from Phonology in the 21st Century: In Honour of Glyne Piggott. McGill Working Papers in Linguistics*, Vol. 22, eds. J Loughran, A McKillen

Pater J, Jesney K, Staubs RD, Smith B. 2012. Learning probabilities over underlying representations. *Proc. Twelfth Meet. Spec. Interest Group Comput. Morphol. Phonol.*, pp. 62–71. Association for Computational Linguistics

Peperkamp S, Le Calvez R, Nadal JP, Dupoux E. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*. 101(3):B31–41

Pierrehumbert JB. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In *Frequency and the Emergence of Linguistic Structure*, eds. JL Bybee, PJ Hopper, pp. 137–157. Amsterdam: John Benjamins

Prickett B. 2018a. Similarity-based Phonological Generalization. *Proc. Soc. Comput. Linguist.* 1(1):193–96

Prickett B. 2018b. Complexity and naturalness biases in phonotactics: Hayes and White (2013) revisited. *Proc. Annu. Meet. Phonol.* 5(0):

Prince A. 2010. *Counting Parses*. Work. Pap., Rutgers University, New Brunswick, NJ

Prince A, Smolensky P. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. John Wiley & Sons

Prince A, Tesar B. 2004. Learning Phonotactic Distributions. In *Fixing Priorities: Constraints in Phonological Acquisition*, eds. R Kager, J Pater, W Zonneveld, pp. 245–91. Cambridge: Cambridge University Press

Rasin E, Berger I, Katzir R. 2015. *Learning rule-based morpho-phonology*. Work. Pap., MIT, Cambridge, MA

Rasin E, Katzir R. to appear. On Evaluation Metrics in Optimality Theory. *Linguist. Inq.*

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. *Struct. Phonol. Represent.* 337–84

Shih SS. 2018. Learning lexical classes from variable phonology. *Proc. AJL2*. 1–15

Smith B. 2015. *Phonologically Conditioned Allomorphy and UR Constraints*. PhD Thesis thesis. University of Massachusetts Amherst

Smith BW, Pater J. 2017. *French schwa and gradient cumulativity*. Work. Pap., University of California Berkeley and University of Massachusetts Amherst

Smolensky P, Goldrick M. 2016. Gradient symbolic representations in grammar: The case of French liaison. *Ms Johns Hopkins Univ. Northwest. Univ.*

Smolensky P, Legendre G. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. MIT press

Soderstrom M, Mathis D, Smolensky P. 2006. Abstract genomic encoding of Universal Grammar in Optimality Theory. In *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, pp. 403–71

Solomonoff RJ. 1964. A formal theory of inductive inference. Parts I and II. *Inf. Control*. 7(1):1–22, 224–254

Stanton J. 2016. Learnability shapes typology: the case of the midpoint pathology. *Language*. 92(4):753–791

Staubs RD. 2014. *Computational modeling of learning biases in stress typology*. PhD Thesis thesis. University of Massachusetts Amherst

Staubs RD, Pater J. 2016. Learning serial constraint-based grammars. In *Harmonic Grammar and Harmonic Serialism*, eds. JJ McCarthy, J Pater. London: Equinox Press

Steriade D. 2001. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. *Ms UCLA*

Tesar B. 1995. *Computational Optimality Theory*

Tesar B. 2006. Faithful Contrastive Features in Learning. *Cogn. Sci.* 30(5):863–903

Tesar B. 2013. *Output-Driven Phonology: Theory and Learning*. Cambridge University Press

Tesar B, Prince A. 2007. Using phonotactics to learn phonological alternations. In *Proceedings of the Thirty-Ninth Conference of the Chicago Linguistics Society*, pp. 209–37. Chicago: Chicago Linguistics Society

Tesar B, Smolensky P. 1998. Learnability in Optimality Theory. *Linguist. Inq.* 29(2):229–68

Tesar B, Smolensky P. 2000. Cambridge, Massachusetts: MIT Press

Tessier A-M-. M. 2009. Frequency of violation and constraint-based phonological learning. *Lingua*. 119(1):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(3):481–87

Wexler K, Culicover P. 1980. 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):1–36

Wilson C. 2006. Learning Phonology With Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cogn. Sci.* 30(5):945–82

Wilson C, Gallagher G. to appear. Accidental gaps and surface-based phonotactic learning: a case study of South Bolivian Quechua. *Linguistic Inquiry*

Yang CD. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press

Yu KM. 2018. Advantages of constituency: computational perspectives on Samoan word prosody. *Proc. 22nd Int. Conf. Form. Gramm.*, pp. 105–124. Springer

Zuraw K. 2000. *Patterned Exceptions in Phonology*

Table 1 - Example of English variable t/d deletion based on Coetzee & Pater (2011)

Input	Output	*CT	MAX-P-V	MAX-FIN	MAX	HARMONY	PROBABILITY
		$w_1 = 4$	$w_2 = 1$	$w_3 = 2$	$w_4 = 3$		
i) /Ct/	[Ct]	–1				$(-1)*w_1 = -4$	≅ 73.1%
	[C_]			–1	–1	$(-1)*w_3+(-1)*w_4 = -5$	≅ 26.9%
ii) /CtC/	[CtC]	–1				$(-1)*w_1 = -4$	≅ 26.9%
	[C_C]				–1	$(-1)*w_4 = -3$	≅ 73.1%
iii) /CtV/	[CtV]	–1				$(-1)*w_1 = -4$	= 50.0%
	[C_V]		–1		–1	$(-1)*w_2+(-1)*w_4 = -4$	= 50.0%

TERMS & DEFINITIONS

1. Hidden Structure

   Any abstract representation that underlies linguistic knowledge but which is not directly observable in the learning data, such as metrical footing, underlying representations, and exceptionality diacritics.

2. 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.

3. 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.

4. Gradient Phonotactics

   Knowledge of legal and likely sound combinations that make up words in a language.

5. Lexical Classes

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

6. Patterned Exceptionality

   When systematic phonological factors statistically condition phonological variation in the aggregate across the lexicon but individual words exhibit fixed behavior.

7. Online Learning Algorithm

   An algorithm that incrementally processes learning data, making updates on a word-by-word basis.

8. Batch Learning Algorithm

   An algorithm that processes the learning data en masse, making updates after consulting the entire data set.

9. 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.

10. 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.

11. 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.

12. Subset Problem

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

13. Bias-Variance Tradeoff

    The balance between tightly fitting observed data (low bias) and generalizing appropriately to unseen data (low variance).

14. Wug Test

    A task (Berko 1958) used to test productivity of morpho-phonological knowledge by asking speakers to produce (or rate) a morphological derivative of a nonce word.

15. Substantive Bias

    A type of inductive, or analytic, bias that favors the learning of patterns with perceptual or articulatory motivations.

16. Analytic Bias

    A cognitive predisposition, or inductive bias, that makes learners more receptive to some patterns than others (Moreton 2008).

17. Channel Bias

    Phonetically systematic errors in language transmission between speaker and hearer (Moreton 2008).

18. 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.

19. Iterated Learning

    A type of agent-based model that simulates vertical transmission of language across generations by modeling ‘parent-child’ interactions where 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.

20. Interactive Learning

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