Abstract. Individual languages exhibit variation in phonological structures both within and across words. This chapter examines the constraints placed on the acquisition of these stochastic phonological patterns. Theoretical perspectives on phonological acquisition are reviewed, revealing the importance of both formal and substantive constraints in a wide variety of learning theories. Empirical studies of the acquisition of variation are then reviewed. These results provide evidence that both formal and substantive constraints are active in the acquisition of variation. The integration of substantive constraints into theories of learning is then discussed, focusing on an extension of Warker and Dell’s (2006) connectionist proposal.

1. Introduction

It is well known that phonological structures vary within as well as across languages and that individual speakers can be sensitive to such variation. A single lexical item can vary due to phonological processes, such as variable word-final /t, d/ deletion in English consonant clusters (e.g., “left”->/lEf/; Guy 1980, et seq.). Across lexical items, variation can be characterized in terms of probabilistic phonotactic constraints. For example, in English both /s/ and /j/ are permitted word initially, although the former is much more frequent (Vitevitch, Armbrüster, and Chu 2004). These stochastic phonological patterns—found within and across words—are a
critical component of our linguistic competence (Labov 1994; but see Newmeyer 2003 for an alternative perspective). More generally, empirical studies suggest that knowledge of stochastic patterns is an integral part of not only linguistic but also other cognitive systems in a variety of species.

How are such patterns encoded by learners? Across domains, the dominant perspective assumes that learners exhibit “probability matching”; learners match their behavior to the patterns present in the environment to the greatest extent possible. This chapter focuses on clarifying the meaning of this final caveat. We review how two types of constraints on learning define the realm of the possible in theories of phonological acquisition. Formal constraints concern the limitations placed on acquisition by the structural properties of the cognitive system. For example, does the cognitive system make use of linear or autosegmental phonological representations? What are the formal operations that define the space of possible rules or constraints of the phonological grammar? In contrast to these structural influences, substantive constraints on acquisition make reference to the intrinsic phonetic properties of phonological representations; they concern not the structure but the content of phonological elements. We then review evidence showing the influence of both of these types of constraints on the acquisition of stochastic patterns. We conclude by discussing the implications of these findings for acquisition theories, illustrating how substantive biases can be incorporated into a connectionist account of the acquisition of stochastic patterns.

2. Constraints in theories of language acquisition

2.1. Probability matching

Probability matching claims that internal cognitive processes are adjusted so as to reflect (or match) the probabilities of the environment. For example, an individual acquiring English
would be exposed to many more strings with /s/ in onset compared to /j/. A probability matching theory of phonological processes in speech production would claim that during the course of learning such processes would be adjusted so that their functioning would reflect these probabilities. For example, in a spreading activation framework (e.g., Dell 1986), /s/-onset may be assigned a higher resting activation level compared to /j/-onset. This encoding of the stochastic pattern would allow the behavior of speakers to be sensitive to this stochastic pattern; e.g., due to its higher activation level, in onset position /s/ would be retrieved more quickly and accurately than /j/. Empirical evidence suggests such patterns can be encoded; English speakers more quickly produce the names of pictures with high frequency as compared to low frequency phonotactics (Vitevitch et al. 2004).

Probability matching is a cognitive interpretation of statistical methodologies in which the parameters of a statistical model are estimated so as to maximize the fit (or minimize the divergence) between the empirically observed data and that generated by the statistical model. For example, many models make use of the method of maximum likelihood estimation, where model parameters are set such that the observed (training) data are assigned the highest possible probability (see Myung 2003, for an introduction; Aldrich 1997, for a historical overview). This is a sensible approach; if a statistical model is to be used to understand a given data set, the model should reflect the properties of the data as closely as possible.

In the phonological domain, many different theories have adopted this general principle as a cognitive claim. For example, connectionist phonological theories (Dell, Juliano and Govindjee 1993; Joanisse 1999; Plaut and Kello 1999; Warker and Dell 2006) assume that phonological structure is generated via spreading activation between simple processing units. Phonological representations are distributed patterns of activity across these units. To generate
variation in such approaches, the relative activation of different output representations can be interpreted as corresponding to relative output probabilities (Warker and Dell 2006). For example, if a /k/ unit is active at 1.0 while /t/ is active at 0.5, the system can be interpreted as producing both /k/ and /t/ (with /k/ being produced with a higher probability). Crucially, learning algorithms for connectionist models are designed such that the model will mimic its training environment as closely as possible. In such theories learning involves modifying the spread of activation so as to minimize the difference (e.g., sum squared error; Kullback-Leibler distance) between the network’s current output and the output specified by the ensemble of training patterns (Rumelhart, Durbin, Golden and Chauvin 1996)—that is, adjusting model parameters so that its outputs match the training environment as closely as possible.

Similar approaches are also found in phonological theories incorporating explicit notions of grammatical structure. Variable rule approaches such as Guy (1991) assume that phonological structures are generated by the ordered application of re-write rules. Stochastic patterns are encoded by allowing rules to be associated with varying probabilities of application. Variation can result if, for instance, a rule applies probabilistically. For example, if a probabilistic cluster simplification rule applies, an underlying cluster will appear as a singleton consonant; if it fails to apply, a cluster will be produced. Crucially, these probabilities are estimated by the theorist directly from the observed frequency of variable processes in the environment (Guy 1991), suggesting that learners assign probabilities by matching the structure of the environment.

Similar assumptions are made in more recent constraint-based formalisms. For example, the Gradual Learning Algorithm (GLA) proposed by Boersma (1997; see also Boersma and Hayes 2001) is a variationist theory stated within the framework of Optimality Theory (OT;
In the GLA, constraints are associated with a probability distribution along an (arbitrary) continuous scale. To produce an output, each constraint is associated with a particular value (a selection point) drawn from this probability distribution. The rank order of these selection points (highest-lowest) is then used to determine a strict ranking of the constraints (highest ranked-lowest ranked). Variation results when two constraints’ probability distributions overlap. In this situation, the rank order of their selection points will switch with non-zero probability—yielding multiple possible constraint rankings and potentially producing variation. Crucially, the GLA uses a variety of stochastic gradient ascent (Jäger 2004) to parameterizes the generation of selections points. This algorithm attempts to match (as closely as possible) the conditional probability of trained outputs given their corresponding input representations (however, see Keller and Asudeh 2003, Pater 2005, for discussion of issues with this algorithm).

An alternative variationist formulation of OT phonological grammars makes use of Maximum Entropy models (Goldwater and Johnson 2003; Wilson 2006). Rather than stochastically generating sets of strict rankings as in the GLA, in Maximum Entropy models the continuous numerical constraint ranking is used to directly generate the probability of various outputs. Specifically, the log probability of an output is proportional to the weighted sum of constraint violations. Variation can therefore result if this weighted sum is relatively close in value for multiple outputs. Critically, the constraint weights are selected by algorithms (e.g., conjugate gradient) that are designed to allow the model to match the structure of the environment (through maximum likelihood methods). This principle can be seen in the name Maximum Entropy; these “are models that have as high an entropy as possible under the
constraint that they *match the training data* [emphasis ours] (Goldwater and Johnson 2003: 112).”

In a variety of frameworks, then, we see that learning is conceptualized as probability matching—adjustment of internal parameters so that behavior matches the statistics of the environment as closely as possible. It is important to consider the implication of the caveat “as closely as possible.” Clearly, learners do not simply mimic the environment; their own internal capabilities influence their ability to reproduce its structure. Learners do not always generalize the patterns they have been exposed to (be they categorical or probabilistic) to unseen data outside the training set. To take an extreme example, no matter how much training data you provide, humans will not be able to generalize (much less produce) a stochastic pattern involving hypersonic frequencies (e.g., producing 20kHz tones with probability 90% and 100kHz tones with probability 10%). Since our vocal tracts are incapable of producing such sounds, there is simply no means by which our behavior could match this aspect of the environment. The following sections consider more subtle aspects of constraints on the ability of speakers to acquire linguistic patterns.

2.2. Formal constraints on learning

All successful learning algorithms make assumptions about the formal nature of the cognitive model that can be used to generate behavior. To construct a processing model, one must assume certain representational and processing mechanisms. These assumptions make predictions regarding what types of patterns individuals can acquire and represent. For example, connectionist networks typically adopt a fixed architecture. The size and structure of input and output representational layers are pre-set, as are the connectivity pattern and size of internal “hidden” representational layers (but see Fahlman and Lebiere 1990 for an alternative approach).
Learning in such networks is limited to the adjustment of weights on connections. These architectural assumptions shape the models that networks adopt to represent the training data. For example, research in the context of reading has shown how the choice of input representations influences network behavior. Seidenberg and McClelland (1989) proposed a connectionist theory of reading where input grapheme and output phoneme representations were context-sensitive triples (e.g., BE is represented by two units: _BE; BE_). Due to this context dependence, the network has no representation that is common across all instances of the grapheme; for example, the B in BE is represented differently than the same letter in BY (i.e., _BE vs. _BY). As shown by Besner, Twilley, McCann and Seergobin (1990) these context-sensitive representations do not support adequate generalization. That is, the networks have great difficulty generalizing the pronunciation patterns present in trained words to untrained nonwords (e.g., “blinch”). As shown by Plaut, McClelland, Seidenberg and Patterson (1996) adopting a less context-dependent representational scheme—where graphemes are not always represented in terms of their neighboring letters—allows the network to develop more structured internal representations that better support generalization. Instead of representing the grapheme B as _BE, in Plaut et al.’s theory, this grapheme is represented simply as Onset-B (such that both BE and BY have overlapping representations). (See also Brousse and Smolensky 1989; Plaut and McClelland 1993 for further discussion of how context-independent or componential representations such as these support generalization in connectionist networks.) Thus, the internal model that connectionist networks acquire is shaped by their representation of the environment.

The impact of formal assumptions on acquisition is much more transparent in grammatical formalisms (e.g., variable rules or OT grammars), which are stated over particular
representational structures and avail themselves of particular formal operations (i.e., particular types of rules or constraints). These restrictions influence the types of phonological patterns—be they stochastic or categorical—that such grammars are capable of encoding. Representational assumptions clearly impact the patterns that grammars can easily describe. Consider the phenomenon of tonal stability, where a tone persists despite the deletion of the vowel on which it was realized. Such a pattern is extremely difficult to capture within linear phonological representations that assume tones to be simply features of vowels. In contrast, it is easily accounted for in a representational framework where tones occupy an independent autosegmental tier (Goldsmith 1976). With respect to formal operations, in OT grammars restricting the types of mechanisms that can be used to define constraints limits the types of grammatical patterns than can be produced. For example, Smolensky (1995, 2006) proposes the formal operation of local conjunction. This defines a complex constraint which is violated only in the case of simultaneous violation of two basic constraints in a specified domain. As discussed by Moreton and Smolensky (2002), by restricting the conjunction to a specified domain this definition limits the types of phonological patterns that OT grammars can describe—namely, blocking the ability to characterize patterns that require sensitivity to violations in two different domains.

Although the role of formal constraints is considerably less transparent for connectionist computational models than generative grammars, their role has been more precisely articulated in other computational frameworks. For example, Gildea and Jurafsky (1996) augment an natural language processing algorithm with a set of formal biases informed by concepts from phonological theory. In a series of simulation studies, they show that these biases allow the
learning algorithm to induce more efficient and general computational devices for encoding phonotactic patterns.

Formal constraints on learning are not limited to hard-and-fast restrictions on the space of possible patterns that can be represented. For example, in Bayesian approaches learning assumptions exert soft pressures on acquisition. In this framework, learning is viewed as the attempt to assign probability to various potential hypotheses given observed data. As shown below, Bayes’ rule allows one to calculate this probability using three values: the probability that if hypothesis X was true, the data would be observed (P(D|H_x)); the probability of observing the data (P(D)); and, most critically for the discussion here, the "prior" probability of the hypothesis X itself (P(H_x)).

\[
P(H_x \mid D) = \frac{P(D \mid H_x)P(H_x)}{P(D)}
\]

In such a framework, we can view categorical restrictions as merely a limit case, where certain hypotheses are assigned zero prior probability. For example, instead of restricting the formal operation of constraint conjunction to single domains, grammars using constraints conjoined across domains could be assigned a prior probability of zero. More complex acquisition theories can be defined when certain hypotheses are assigned prior probabilities that are merely lower (not simply zero). A common approach is to formally instantiate a version of Ockham’s Razor\(^1\) by assigning low (but non-zero) probabilities to more complex hypotheses. For example, Goldwater and Johnson (2003) applied a Bayesian perspective to learning constraint weights in a Maximum Entropy model. Following Johnson, Geman, Canon, Chi and Riezler (1999), they assumed the prior probability for each weight is a Gaussian with a mean of zero—biasing constraint weights to be as small as possible given the training data. Since in a

\[^1\]"Entities should not be multiplied unnecessarily."
Maximum Entropy framework constraints with zero weight do not influence the output of the grammar (they do not contribute to the sum of weighted constraint violations), this is equivalent to preferring less complex hypotheses with fewer active constraints.

Moreton (2006) proposes a different Ockham-inspired restriction on the learning of constraint weights. He notes that OT theories incorporating constraints that are “modular” in the sense that they refer to only a single level of structure (e.g., tones only; segments only) allow for a smaller number of possible grammars than theories incorporating constraints that refer to multiple levels of structure simultaneously. He then proposes that learners acquire OT grammars by using a Bayesian algorithm to gradually incorporate all types of constraints. Assuming a uniform prior over grammars, the prior probability of a particular grammar that incorporates a constraint greatly expanding the hypothesis space will be lower than that of a particular grammar that does not expand the space to the same degree. In the former case, due to the uniform prior over grammars the large number of competing hypothesis will eat up too much of the probability mass, lowering the probability of any particular grammar. In Moreton’s Bayesian framework, this lower prior probability leads learners to favor grammars that use modular constraints. However, in the presence of sufficient data, they can acquire non-modular grammars as well, illustrating the formal constraints on learning can exert soft pressures on acquisition.

2.3 Substantive constraints on learning

The theoretical claims discussed above make reference only to the formal properties of mechanisms that represent linguistic structure. For example, in a Maximum Entropy context Ockham’s Razor prefers grammars that have fewer constraints—it does not matter what the constraints refer to, simply that they be small in number. In contrast, other theories have assumed that learning mechanisms are in addition explicitly and directly constrained by the
substance of phonological patterns. Such theories claim that because such constraints are present within learning mechanisms patterns motivated by perceptual, articulatory or cognitive factors are more easily acquired than those that are not.2

Many grammatical theories have explicitly incorporated such constraints in mechanisms that define grammatical rules or constraints. In derivational formalisms, substantive considerations have been seen as evaluation criteria for grammatical analyses (i.e., ranking certain rules or grammars that contain phonetically unmotivated patterns as dispreferred; Chomsky and Halle 1968, Archangeli and Pulleyblank 1994). More recently, Hayes (1999) directly incorporated such constraints into the acquisition process, proposing that learners use substantive information when inducing OT constraints. Formal constraints on this acquisition process are as in other OT frameworks; constraints are limited by the representational and formal properties of the grammatical theory. In addition, Hayes incorporates a phonetic map specifying the relative perceptual/articulatory difficulty of output forms. This map is used by learners to filter the set of possible constraints. If a constraint’s violation profile fits the phonetic map better than other formally similar constraints (i.e., by assigning violations to difficult to perceive rather than easily perceived forms with higher probability than other constraints), it is admitted to the set of constraints entertained by the learner. This filtering process restricts the types of

2 Others have claimed that the influence of substantive factors arises as a byproduct of diachronic processes (e.g., misperceptions) and does not reflect the direct encoding of substantive factors by processes that represent and acquire phonological patterns (e.g., Blevins 2004, Hale and Reiss 2000, Ohala 1981). Such proposals have difficulty accounting for results where learners show sensitivity to substantive factors in the face of unbiased training data. For example, the Goldrick and Larson (2007) study discussed in section 3 shows substantively-motivated differences in the acquisition of phonotactic patterns. This occurs even though participants receiving equivalent training data across conditions (see below for further discussion). Since the training data provide no evidence of substantive factors, these factors must be directly influencing acquisition processes (see Wilson 2006 for similar results and further discussion).
phonological patterns that OT grammars can represent—namely, to those that respect to at least some degree the substantive properties that phonetic map encodes.

Just as with formal constraints on learning, substantive constraints are not limited to categorical, hard and fast restrictions on phonological patterns; in a Bayesian framework, grammars incorporating phonetically dispreferred constraints can be assigned merely lower (but non-zero) prior probabilities. For example, Wilson (2006) presents a Maximum Entropy framework for OT grammars where the prior on constraint weights is not constant across all formally equivalent constraints. The priors are structured so that constraints that can trigger a highly phonetically motivated change are more likely to be active than those that trigger less phonetically motivated processes (e.g., changes that result in highly perceptually similar forms are favored over changes that result in less perceptually similar forms). Learners are therefore more likely to adopt grammars that incorporate phonetically favored vs. disfavored constraints (although it is possible for both types of grammars to be acquired).

3. Empirical evidence for constraints on acquisition

As discussed above, most theories (implicitly or explicitly) adopt the claim that learners match the probabilities in the environment. This is supported by sociolinguistic research showing that learners can and do acquire stochastic patterns present in their environment (e.g., Roberts 1997; see Labov 1994, for a review). As the preceding sections have shown, however, theories also assume that there are important limitations on this ability; learners probability match to the greatest extent possible. Furthermore, different theories make quite different claims regarding what is possible for learners to acquire, incorporating a variety of both formal and substantive constraints on acquisition processes. In the following sections, we review evidence that these two types of constraints influence the acquisition of stochastic patterns specifically.
Following a variety of stochastic formalisms discussed above, the following discussion assumes that a common set of mechanisms supports the acquisition and representation of both stochastic and categorical patterns. Categorical patterns are not seen as qualitatively different from stochastic patterns; categorical patterns simply represent a limit case, where one outcome has a probability near 100%. For example, in a variable rule formalism, a rule can be associated with a 100% probability of application. In a GLA or Maximum Entropy-based OT formalism, relevant constraint weights can be so widely separated as to essentially eliminate the probability that an alternate form will be produced. The discussion is therefore built on the premise that the formal and substantive biases discussed below are not biases for using completely distinct categorical vs. stochastic learning/representational systems; they represent constraints on the learning/representation of internal parameters of a single system representing both categorical and stochastic patterns.

3.1 A general dispreference for stochastic patterns

Studies across multiple domains suggest that stochastic patterns can be difficult for learners to acquire. In many situations, learners do not acquire and generalize a stochastic pattern. Instead, they regularize; learners categorically apply a single variant to their productions, disregarding variation in the training data. Since this result is observed across multiple domains, it is consistent with a constraint on acquisition processes that simply disfavors variability. This constraint is formal in that it disfavors all stochastic patterns; it does not distinguish patterns based on their substantive properties.

Elissa Newport and her colleagues have documented a number of cases of regularization in learning at both morphological and syntactic levels of structure. Singleton and Newport (2005) reported a case study of Simon, a child learning American Sign Language (ASL).
Outside of the experimenters involved in the study (who only briefly interacted with Simon on occasions separated by several months), Simon’s ASL input came exclusively from his parents. His parents had acquired ASL at a late age. Like many late acquirers, their usage patterns were much more variable than native ASL signers. For example, they tended to omit morphemes or replace these morphemes with periphrastic constructions. Interestingly, in the face of this variable, inconsistent input, Simon rarely omitted morphemes, in many cases using them at rates well within the range of children exposed to native ASL signing input. Thus, even though Simon’s input was the variable productions of his parents, his productions were near-categorical.

With respect to syntactic structure, in another series of studies Newport and colleagues have examined the acquisition of word order in artificial grammar learning tasks. Hudson Kam and Newport (2005) examined the ability of adults and children to learn a novel language in which determiners could appear in either pre- or post-nominal position (e.g., English glosses would read “the dog” vs. “dog the”). To expose participants to stochastic patterns, the rate at which a determiner was present in a specified position was manipulated. For example, in one condition 60% of the training sentences contained phrases like “the dog” while the other 40% contained phrases like “dog” with no determiner. Under certain circumstances, probability matching was observed. When adults were tested specifically on nouns that had appeared in the training sentences, their behavior generally matched these stochastic patterns (Hudson Kam and Newport 2005). For example, when asked to describe scenes containing objects that had appeared in the training sentences, their rate of determiner use matched the rate at which determiners were present in the training sentences. However, in other situations, participants failed to acquire the stochastic patterns. In this same paradigm, Hudson Kam and Newport found that children were more likely to adopt a categorical approach. When 60% of the training
sentences contained a determiner, only 29% the children adopted a variable production strategy; the remaining 71% either always produced determiners or never produced determiners. This suggests that child language acquisition is constrained in a way that disfavors stochastic syntactic patterns. Wonnacott and Newport (2005) found evidence that adult learning is also subject to similar constraints; adults failed to generalize a stochastic pattern to novel sentences. When adults were trained on either variable noun-determiner orders (as above) or variable object-subject orders (i.e., verb-object-subject vs. verb-subject-object), their behavior matched the stochastic patterns only when they were tested on nouns present in the training sentences. When tested on nouns they had seen only in a simple vocabulary test (and never in a sentence context), the majority of participants applied a single pattern to all forms (e.g., they produced all objects before subjects or vice versa; the choice of which pattern was to a certain degree idiosyncratic across participants). This suggests that like children, adult acquisition is also subject to constraints that disfavor variation; when processing novel forms, adults appear to regularize, applying only a single pattern present in the training data.

Failure to generalize stochastic patterns to novel items has also been observed in non-linguistic domains. A well-studied problem from the decision-making literature concerns learning in the context of binary choice problems with limited initial information and immediate feedback. For example, a participant is asked to predict which of two lights will be turned on. After each prediction the participant receives a reward if they are correct (and loses or fails to gain credit based on incorrect predictions). When confronted with a stochastic pattern (e.g., light 1 yields a reward 60% of the time), participants’ choices often match the probability of rewards (e.g., light 1 is guessed 60% of the time; Shanks, Tunney and McCarthy 2002). However, under certain circumstances individuals will regularize, consistently picking the outcome that yields the
largest reward\textsuperscript{3}. For example, Shanks et al. (2002) found individuals could learn to pick the highest yield outcome when financial rewards were large, explicit feedback was regularly provided, and individuals received extensive training (see also Erev and Barron, 2005, for an extensive review of similar situations where probability matching is not observed). Studies of learning in binary choice tasks in other species show a similar mixture of results. When presented with multiple food sources (where food is distributed according to some stochastic pattern) animals often exhibit probability matching. The probability that they will visit a food source first is roughly equal to the probability that food will be located there. For example, if one arm of T-shaped maze has food 80% of the time, the animal will go down that arm first on 80% of trials. However, when animals are prevented from visiting the multiple food sources on each trial (e.g., after visiting one food source, they can visit no other sources), they consistently choose the source with the highest probability of food. Following the example above, if the experimenter instead closes off one arm of the T-maze after the animal goes down the other, the animal will consistently choose to travel down the arm where food is located on 80% of trials (see Gallistel and Gibbon 2002 for a review).

These studies show that although humans and other animals can under certain circumstances match the probability structure of their environment, they are not bound to do so. When exposed to stochastic patterns, they sometimes regularize, producing categorical behavior. This is consistent with a constraint on learning processes that generally disprefer stochastic patterns.

\textsuperscript{3} Note that this pattern of behavior maximizes the expected reward participants will receive, making it the rational strategy for participants to adopt (see, e.g., Shanks et al. 2002 for discussion). It is therefore possible that participants have in fact acquired the probabilistic distribution of outcomes and are rationally using this information as a basis for categorical behavior.
3.2 A dispreference for specific stochastic patterns

Formal constraints on acquisition predict general restrictions on all formally similar phonological patterns (e.g., a general dispreference for stochastic patterns). In contrast, substantive constraints predict that formally similar patterns may be acquired differently. In this section, we briefly review our recent work (Goldrick and Larson in press) that examines this question. We tested for the presence of substantive constraints on acquisition by examining the ability of participants to implicitly acquire a variety of formally similar stochastic patterns. Specifically, we examined the acquisition of phonotactic patterns where a single segment was restricted to a particular syllabic position (e.g., /f/ is restricted to onset). All these patterns therefore have the same “structural” complexity; for example, the description of each pattern refers to the same number of autosegments. In the absence of structural or formal distinctions among the patterns, differences in participant’s ability to learn them can therefore only be attributed to the substance of the patterns—that is, the actual phonetic content the patterns are stated over.

To examine acquisition, we utilized an implicit learning technique developed by Dell, Reed, Adams and Meyer (2000; see also Dell and Warker 2006; Goldrick 2004; Taylor and Houghton 2005). This technique relies on two features of speech error patterns. First, speech errors tend to respect syllable position (see Vousden, Brown, and Harley, 2000, for a recent review). If there is an error in which a segment moves from its target position and either replaces or switches with another segment, it tends to move from one particular position in its word to the same position in another word. An onset is therefore more likely to move to another onset, and a coda to another coda. For example, given a target sequence like “beer can,” an error like “beer ban” (onset -> onset) is more likely than an error like “beer cab” (onset -> coda). The
second critical feature of speech errors is that this tendency to respect syllable position is massively strengthened when segments are subject to phonotactic constraints (see Vousden et al. 2000, for a recent review). For example, /h/ is confined to onset position in English lexical items. Given a target sequence like “hear can,” an error like “hear han” (onset -> onset) is much, much more likely than an error like “beer cab” (onset -> coda; the latter type of error is observed extremely infrequently).

Dell et al. (2000) used these two features of speech errors to examine the influence of novel phonotactic constraints that restricted English consonants to particular syllable positions. English-speaking participants were asked to read aloud tongue twisters composed of four CVC syllables. As in English, some consonants in these syllables freely occurred in both onset and coda (e.g., /m/ appeared in both onset and coda with equal frequency). In addition, a pair of critical consonants reflected constraints specific to the experiment (e.g., /f/ occurred only in onset). Participants read aloud these tongue twister three times quickly to induce speech errors. (There was therefore no explicit “training” session; participants simply repeated tongue twisters exhibiting the phonotactic patterns of the experiment.) As in previous studies, participants’ speech errors tended to respect syllable position (e.g., when /m/ was intended to be produced in onset, it tended to surface in onset in errors). Interestingly, following naturally occurring phonotactic constraints, this effect was massively strengthened by the constraints specific to the experimental syllables. For example, when /f/ was restricted to onset, only 3% of the erroneously produced /f/ segments occurred in coda. When producing a sequence like “fem nep,” only 3% of participants’ errors were of the form “fem nef”; 97% were of the form “fem fep.” This result was uninfluenced by whether or not participants were aware of the phonotactic
restriction, suggesting that in this paradigm speakers can implicitly acquire new phonotactic constraints.

Following this work, we examined the acquisition of phonotactic patterns associating particular segments to onset or coda position. We extended this work by exposing participants to stochastic patterns. One of two critical consonants (/f/ or /s/, following previous studies with this paradigm) was associated to onset or coda position with varying probabilities, as shown in Table 1. For example, in the first condition /f/ appeared in onset in 20% of the tongue twister sequences (e.g., “feng nep kem gez”) and in coda in the remaining 80% (e.g., “hef nep kem gez”). The critical consonants were combined with a set of control segments (all more than 1 feature distinct from the critical consonant) to form CVC syllables. Following Dell et al., these were embedded in tongue twisters and the distribution of participants’ speech errors were used to index the influence of newly acquired phonotactic constraints.

Table 1. Distribution of critical consonant in onset vs. coda across all tongue twisters in each condition (Goldrick and Larson in press).

<table>
<thead>
<tr>
<th>Percent of tongue twisters where critical segment is in onset</th>
<th>20%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of tongue twisters where critical segment is in coda</td>
<td>80%</td>
<td>60%</td>
<td>50%</td>
<td>40%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Results are summarized in Figure 1, separated by each critical segment (/f/ vs. /s/) as well the prosodic position the segment was intended to be produced in (onset vs. coda). (For full details and statistical analyses of the results, see Goldrick and Larson in press.) For example, the line with filled squares shows results for tongue twister sequences where /f/ was intended to be
produced in onset position (e.g., “feng nep kem gez”). The x-axis details the proportion of all tongue twister targets in which the specified segment appeared in that syllable position. For example, the leftmost point in the /f/ onset line characterizes data from the condition in which the target position for /f/ was onset in 20% of sequences (e.g., “feng nep kem gez”) and coda in 80% of sequences (e.g., “hef nep kem gez”). Finally, the y-axis shows the mean proportion of participant errors resulting in the production of the critical segment that appeared in that syllable position. For example, the leftmost point in the /f/ onset line shows the mean proportion of errors resulting in the production of an /f/ that appeared in onset position (e.g., the mean proportion of errors of the type “feng nep”->“feng fep” as opposed to “feng nep” -> “feng nep”).

Figure 1. Results of Goldrick and Larson (in press). The lines separate results by critical segment and syllable position. The x-axis shows the proportion of sequences in which the critical segment appeared in that syllable position. The y-axis depicts the mean proportion of errors across participants resulting in that critical segment in that syllable position.
For many of the stochastic patterns, the pattern of performance was clearly influenced by
the probabilistic phonotactics of the tongue twister sequences (although there was considerable
variance; we return to this below). The proportion of participants’ errors resulting in /f/ in onset
or /f/ in coda generally varied according to the frequency with which /f/ appeared in that syllable
position in the tongue twister targets. For example, as shown by the line marked by open squares
in Figure 1, participants were more likely to produce an error where /f/ appeared in onset
position (e.g., “feng nep”->“feng fep”) when /f/ appeared in onset position in 60% vs. 50% of
tongue twister targets. Similar results were found for /s/ when it was intended to be produced in
coda position; if /s/ appeared in coda in fewer tongue twister targets, participants were in general
less likely to produce errors resulting in /s/ in coda position (e.g., the relative probability of “hes
nep” -> “hes neg” generally went up as the number of /s/-coda targets increased). A quite
different pattern was observed for sequences where /s/ was intended to be produced in onset
position. As shown in Figure 1, errors of the form “seng nep” -> “seng sep” occurred with very
high probability, even when relatively few tongue twister targets had /s/ in onset position.

It is important to note there was a substantial degree of variation in participants’ error
patterns, partially obscuring the influence of probabilistic phonotactics. For example, for /s/ in
coda, the mean proportion of errors resulting in /s/ coda is lower, not higher, in the 50% vs. 40%
condition. A probability matching hypothesis would predict the opposite pattern—if /s/ occurs in
coda in more targets in the 50% vs. 40% condition, errors in that condition should more
frequently result in /s/ coda. There are two points to note here. First, this difference is most
likely due to high degrees of participant variability (i.e., variance of the sample proportion was
24% in the 50% condition). Second, although behavior in the /s/-coda and /f/-onset/-coda
conditions may not precisely reflect the stochastic patterns present in the input, it is clear that
performance in the /s/-onset condition deviates to a far greater degree from the predictions of probability matching.

Why would /s/ onset behave differently from the other conditions considered here? It is well known that /s/ is perceptually salient. For example, relative to other fricatives, it exhibits a greater amplitude of frication noise and a more distinct spectral envelope (see Davidson, 2003, for a review). With respect to onset position specifically, /s/ is quite typologically frequent as the initial member of a cluster; Morelli (1999) in fact argues that it is the least marked cluster type. Finally, prevocalic /s/ is relatively resistant to diachronic lenition processes. As reviewed by Ferguson (1990), /s/ weakening processes (e.g., /s/>/h/) target either intervocalic or pre-consonantal/coda /s/ before generalizing to the prevocalic/onset position.

The distinct pattern of behavior for /s/ onset targets might therefore be accounted for by the intrinsic phonetic properties of the segment involved in this phonotactic pattern. This stands in marked contrast to the studies reviewed in the previous section. There, constraints on acquisition targeted an entire class of formally related patterns (e.g., all stochastic patterns). These results show that constraints on acquisition can distinguish among members of such a class. Participants were exposed to a set of phonotactic patterns involving the probabilistic restriction of a segment to two syllable positions. Formally, these patterns are identical; they are of identical complexity and concern the same representational levels. The only distinction is substantive. To produce the differential patterns of learning performance, constraints on acquisition processes must therefore make reference not just to the formal complexity of phonological patterns but also to the intrinsic phonetic properties of the structures involved.
4. Incorporating substantive constraints in theories of the acquisition of variation

The evidence reviewed above suggests there may be a role for both formal and substantive constraints in the acquisition of variation. As reviewed in section 2, all theories incorporate the former to some extent; however, fewer theories have focused on the additional impact of substantive constraints. A notable exception is that of Wilson (2006). As reviewed in section 2, Wilson’s Maximum Entropy theory makes use of both formal and substantive constraints, biasing the learning algorithm in favor of constraint-based grammars that are both simple and incorporate phonetically motivated constraints. This proposal is therefore most consistent with the existing data. However, it is possible that other probabilistic frameworks such as variable rules (e.g., Guy 1991) or connectionist networks (e.g., Warker and Dell 2006) could be extended to incorporate these results. In this section, we focus on the latter, sketching an extension to Warker and Dell’s (2006) proposal that incorporates substantive constraints within acquisition mechanisms.

Warker and Dell (2006) present a connectionist theory of metrical encoding, a component of spoken production processing where the melodic content of an utterance (e.g., segments) is associated with metrical structure (e.g., syllables). Following other proposals (e.g., Dell 1986; Levelt, Roelofs and Meyer 1999), they assume that this process receives as input a set of segments in a specified order; the output is a representation in which these segments have been associated to syllable positions. They instantiated this process in a three layer neural network which was trained using the backpropogation algorithm. As noted above, to generate probabilistic behavior from the network, they assumed that the relative activation level of various output units reflected the relative probability of responses (e.g., if a unit was highly active, it had...
a high probability of being produced; units with lower activity had lower, but non-zero, probabilities).

Goldrick (in press) considered a simplified version of Warker and Dell’s (2006) network which permits a closed-form analysis. His results show that for this simplified network the activation of segments in particular syllable positions (i.e., output units of the network) reflects the weighted interaction of three factors:

1. Pressure to preserve the intended syllable position of the segment
2. Overall frequency within the training set of the segment in the intended syllable position
3. Overall frequency within the training set of the segment in the non-intended syllable position.

As discussed in Goldrick (in press), the presence of the first factor allows the network to account for the tendency of speech errors to preserve target syllable position (Vousden et al. 2000). Allowing the second and third factors to influence output activation allows the network to capture the influence of newly acquired phonotactic constraints on speech errors (e.g., Dell et al. 2000).

We propose that this theory be augmented by the addition of factors reflecting substantive biases. In a more realistic instantiation, this would be a complex, context-sensitive set of factors incorporating both perceptual and articulatory dimensions of phonetic structure. For the purposes of the simple sketch of an approach offered here, we can use a single simple factor as a “placeholder” for a more richly articulated theory of substantive biases:

4. Phonetic difficulty of the segment in the specified syllable position.
Computationally, this simple factor can be instantiated as a bias on the activation of each output unit. Biases drive the activation of a unit higher or lower regardless of the input. Crucially we assume that these biases do not reflect the distributional properties of the output units (i.e., they are not learned by the same mechanisms that allow the network to be sensitive to the frequency of each syllable position). Instead, these biases are set according to the intrinsic phonetic properties of each segment in each syllable position. They therefore roughly correspond to the values in the phonetic map of Hayes (1999), which specified the relative perceptual/articulatory difficulty of output forms (see section 2 for discussion).

Incorporating these substantive factors into this connectionist theory has implications for its ability to acquire stochastic patterns. If the bias factor is sufficiently strong, the frequency factors will be unable to influence performance. In other words, these substantive factors will constrain acquisition of stochastic patterns present in the environment. To illustrate this influence, consider the Goldrick and Larson (in press) data reviewed in section 3. Assume there is a strong bias for /s/ in onset position. Due to this strong bias, manipulation of the frequency of /s/ in coda will not influence the distribution of errors for /s/ onset. In contrast, when /s/ is intended to be produced in coda position, only the pressure to preserve target syllable position blocks the strong influence of the bias. As soon as the frequency of /s/ coda decreases (and that of /s/ onset increases), the combined influence of the bias and frequency overwhelms this factor, leading to probability matching.

To illustrate these effects, we calculated the predictions of a simple version of this augmented theory following Goldrick (2007). We assumed that output unit activation reflected the linear sum of the factors listed above. The four factors were instantiated in the following way:
1. **Pressure to preserve target syllable position.** The activation of the intended syllable position was increased by 1.0 unit; the activation of the unintended position was decreased by the same amount.

2. **Overall frequency of intended position.** The activation of the intended syllable position was increased by a proportion of 2.0 units; this proportion was fixed by the relative frequency of the intended position (e.g., at a relative frequency of 50%, the second factor contributed 1.0 units). The activation of the unintended position was decreased by the same amount.

3. **Overall frequency of unintended position.** This was the inverse of the second factor above. Here, the activation of the unintended syllable position was increased by a proportion of 2.0 activation units (based on the relative frequency of the unintended position). The activation of the intended position was decreased by the same proportion of 2.0 units of activation.

4. **Phonetic bias.** We implemented 3 different bias scenarios.
   
   a. **Bias favors the target syllable position.** Activation of the intended syllable position was increased by 1.0, while the unintended position was decreased by the same amount. This corresponds to the intended /s/ onset condition above.

   b. **Bias favors the unintended position.** Activation of the intended syllable was decreased by 1.0, while the unintended position was increased by the same amount. This corresponds to the intended /s/ coda condition above.
c. **Unbiased.** Both the intended and unintended positions had a small positive bias (+0.5). This corresponds to the /f/ onset and coda conditions above.

The activation of the segment in the intended (target) position and the unintended syllable position was therefore characterized by the following set of equations:

\[
\begin{align*}
    a(\sigma_{\text{intended}}) &= 1.0 + 2.0\alpha - 2.0(1 - \alpha) + \beta_{\text{intended}} \\
    a(\sigma_{\text{unintended}}) &= -1.0 - 2.0\alpha + 2.0(1 - \alpha) + \beta_{\text{unintended}}
\end{align*}
\]

Where \(\alpha\) is the proportion of sequence where the segment appears in the intended position, and \(\beta\) represents the phonetic bias of the segment in each position (across 3 scenarios here, intended/unintended = +1.0/–1.0; –1.0/+1.0; +0.5/+0.5).

To generate predicted error distributions, we assumed that the activation values reflect the relative well-formedness of each output representation. Drawing on Maximum Entropy models, we converted these relative well-formedness scores to production probabilities using the following formula.

\[
\Pr(x) = \frac{e^{a(x)}}{\sum_{j \in O} e^{a(j)}}
\]

Where \(x\) is an output unit, \(a(x)\) is the activation of output unit \(x\), and \(O\) the set of all output units. In other words, the log probability of a particular output is a function of its relative activation value. Here, we consider only the two output units (the intended target and unintended syllable position). Figure 2 shows the predicted probabilities for the three scenarios outlined above.
Figure 2. Predictions of the connectionist theory augmented by substantive constraints. Biases either favor the intended syllable position (as in the /s/ onset condition above), the unintended syllable position (as in /s/ coda) or neither (as in /f/ onset and coda).

When the system is unbiased, errors are predicted to be sensitive to the stochastic patterns present in the input. As shown in Figure 2, the lower the frequency of the intended syllable position, the smaller the proportion of errors appearing in that position. This tendency is even more strongly observed when substantive biases favor the unintended syllable position. In this situation, the tendency to preserve target syllable position is easily overwhelmed by the influence of frequency-dependent factors. Since the bias favors the unintended position, the pressure to preserve target syllable position is quite weak. In contrast, when substantive biases favor the intended syllable position, frequency manipulations are unable to exert a strong influence. As shown in figure 2, the predicted proportion for this bias scenario is quite close to ceiling at a wide range of frequency values. In the face of both the pressure to preserve target syllable position and the bias towards the same position, the frequency dependent factors are unable to
exert a strong influence on error distributions. By augmenting Warker and Dell’s (2006) proposal to include substantive constraints, this analysis shows that connectionist approaches can also provide a viable account of the influence substantive biases. More generally, by incorporating substantively motivated factors into probabilistic models, learning theories can account for the influence of both formal and substantive constraints on the acquisition of stochastic patterns.

5. Conclusion

The simplicity of the phrase “probability matching” belies the considerable complexity of theories of learning. Learners attempt to match the structure of the environment in the context of both formal and substantive constraints on their learning mechanisms. The research reviewed here shows that the influence of these constraints on acquisition is not limited to the learning of categorical patterns; stochastic pattern acquisition is also subject to both formal and substantive constraints. The challenge for future theoretical work is to elucidate the specific nature of these constraints and as well as integrate them into probabilistic learning models.

References


Goldrick, Matthew and Meredith Larson in press Phonotactic probability influences speech production. *Cognition."


Pater, Joe 2005 Non-convergence in the GLA and variation in the CDA. Unpublished manuscript, University of Massachusetts, Amherst. Rutgers Optimality Archive ROA-780, http://roa.rutgers.edu


