A model of rapid phonotactic generalization

Tal Linzen
Department of Linguistics
New York University
linzen@nyu.edu

Timothy J. O'Donnell
Brain and Cognitive Sciences
Massachusetts Institute of Technology
timod@mit.edu

Abstract

The phonotactics of a language describes the ways in which the sounds of the language combine to form possible morphemes and words. Humans can learn phonotactic patterns at the level of abstract classes, generalizing across sounds (e.g., "words can end in a voiced stop"). Moreover, they rapidly acquire these generalizations, even before they acquire sound-specific patterns. We present a probabilistic model intended to capture this early-abstraction phenomenon. The model represents both abstract and concrete generalizations in its hypothesis space from the outset of learning. This—combined with a parsimony bias in favor of compact descriptions of the input data—leads the model to favor rapid abstraction in a way similar to human learners.

1 Introduction

Natural languages place restrictions on the ways in which sounds can combine to form words (the *phonotactics* of the language). The velar nasal [ŋ], for example, occurs at the end of English syllables, as in *ring* [ɹɪŋ] or *finger* [fɪŋgəɹ], but never at the beginning of a syllable: English does not have words like **ngir* [ŋɪɹ]. English speakers are aware of this constraint, and judge forms that start with a [ŋ] as impossible English words.

Sounds that share articulatory and/or perceptual properties often have similar phonotactic distributions. German, for example, allows voiced obstruents,[1] such as [b] and [g], to occur anywhere in the word except at its end: [bal] is a valid German word, but [lab] isn't.

Speakers use such features of sounds to form phonotactic generalizations, which can then apply to sounds that do not appear in their language. Although no English words start with either [sɹ] or [mb], English speakers judge *srip* to be a better potential word of English than *mbip* (Scholes, 1966); this is likely because [sɹ] shares properties with strident-liquid clusters that do exist in English, such as [sl] as in *slip* and [ʃɹ] as in *shrewd*, whereas [mb] does not benefit from any sonorant-stop onset sequences (*[nt])—none exist in English.

Recent studies have investigated how humans acquire generalizations over phonological classes in an artificial language paradigm (Linzen and Gallagher, 2014; Linzen and Gallagher, 2015). The central finding of these studies was that participants rapidly learned abstract phonotactic constraints and exhibited evidence of generalizations over classes of sounds before evidence of phoneme-specific knowledge.

This paper presents a probabilistic generative model, the Phonotactics As Infinite Mixture (PAIM) model, which exhibits similar behavior. This behavior arises from the combination of two factors: the early availability of abstract phonological classes in the learner's hypothesis space; and a parsimony bias implemented as a Dirichlet process mixture, which favors descriptions of the data using a single pattern over ones that make reference to multiple specific patterns.

2 Summary of behavioral data

The experiments are described in detail in Linzen and Gallagher (2014) and Linzen and Gallagher (2015); we summarize the main details here.

Design: Participants were exposed to varying numbers of auditorily-presented words in one of two artificial languages, VOICING and IDENTITY.

[1] See Hayes (2011) for an introduction to phonological features.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1138–1143,
Lisbon, Portugal, 17-21 September 2015.

Exposure		Test		
ganu	gimi	CONF-	CONF-	NONCONF-
balu	bini	ATT	UNATT	UNATT
vimu	voni	zonu	dila	tumu
zalu	zili	zini	dimu	talu
ðano	ðamu			

Table 1: VOICING: Design for one of the lists (voiced exposure, [d] held out). The table shows the complete list of exposure words that a participant in the two exposure sets group might receive.

Exposure	Test	
	CONF-	CONF-
	ATT	UNATT
pipa	papi	keku
ʃuʃe	ʃeʃu	sasi
gapu	gugi	dʒidʒe
nuni	nanu	mamu
	NONCONF-	NONCONF-
	ATT	UNATT
kesa	kasi	pina
mudʒe	medʒa	nage
dʒuki	dʒuke	gaʃe
semi	sami	ʃipu

Table 2: IDENTITY: A complete list of exposure and test words that a participant in the one exposure set group might receive.

Following the exposure phase, they were asked to judge for a set of novel test words whether those words could be part of the language they had learned (the possible answers were "yes" or "no").

In the VOICING experiment, all exposure words began with consonants that had the same value for their voicing feature (all voiced or all voiceless, e.g., $A_1 = \{$g, v, z, ð, b$\}$ or $A_2 = \{$k, f, s, θ, p$\}$). Some of the sounds with the relevant voicing were held out to be used during testing (e.g., [d] for A_1 or [t] for A_2). All exposure and test words had the form CVMV (*tumi*), where C stands for the onset consonant, V for a randomized vowel, and M for [m], [n] or [l] (see Table 1).

Participants judged three types of novel test words: ones with the same onset as one or more of the words in exposure (CONF-ATT;[2] e.g., *zonu* for A_1); ones whose onset was not encountered in exposure but had the same voicing as the exposure onsets (CONF-UNATT; e.g., *dila*); and ones whose onset had different voicing from the exposure words (NONCONF-UNATT, e.g., *tomu*). The vowels and the second consonant were randomized across conditions such that only the onsets reliably discriminated the three conditions.

In the IDENTITY language, words had the form C_1VC_2V. The generalization in this language was $C_1 = C_2$ (e.g., *pipa*). Here it was probabilistic: only half of the words in the exposure stage conformed to the generalization. As such, there was a fourth test condition, NONCONF-ATT, of exposure words that did not conform to the generalization.

Participants were recruited on Amazon Mechanical Turk (280 participants in the VOICING experiment and 288 in the IDENTITY experiment). They were divided into four groups, which received 1, 2, 4 or 8 sets of words. In the VOICING experiment, each of the sets contained five words, one starting with each of the five CONF-ATT onsets; in the IDENTITY experiment, each of the sets contained eight words, one with each of the CONF-ATT and NONCONF-ATT consonant pairs (Tables 1 and 2).

Results: Human experimental results are plotted in Figure 1. Endorsement rates represent the proportion of trials in which participants judged the word to be well-formed. Participants learned generalizations involving abstract classes of sounds after a single exposure set: in the VOICING experiment, they judged voiced word onsets to be better than voiceless ones, and in the IDENTITY experiment they judged words with identical consonants as better than words with nonidentical ones.[3] Participants did not start distinguishing CONF-ATT from CONF-UNATT patterns until they received two or more sets of exposure to the language.

Participants continued to generalize to CONF-UNATT patterns even after significant exposure to the language. Endorsement rates were higher than 50% across the board, likely because even words with NONCONF-UNATT consonant patterns were similar to the exposure words in all other respects (e.g., length, syllable structure, number of vowels

[2] ATT (attested): the consonants in the word (though not the full word) were encountered in exposure; UNATT (unattested): consonants were not encountered in exposures; CONF (conforming): the consonants conform to the abstract pattern (voicing or identity); NONCONF (nonconforming): consonants don't conform to the abstract pattern.

[3] All differences discussed in this section are statistically significant.

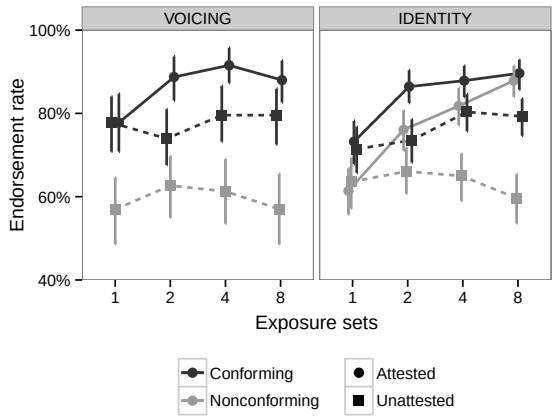

Figure 1: Human behavioral results. Error bars indicate bootstrapped 95% confidence intervals.

and consonants).

3 The model

PAIM is a generative model: it describes the probabilistic process by which the phonological forms of words are generated. Phonotactic knowledge is expressed as a set of word-form templates, represented as sequences of phonological classes. For example, the template ⟨[+voiced], V, C, V⟩ captures a generalization over words beginning with a voiced consonant.

Prior over phonological classes: Phonemes are represented as phonological feature-value matrices.[4] We generate a phonological class for each position in the template using this feature system and a parameter $p \in [0, 1]$, which controls the model's willingness to consider more or less abstract phonological classes: low values of p encourage underspecified classes, such as [] or [+voice], whereas high values of p favor highly specified classes, such as [+voice, labial, -continuant]. Given a particular value of p, we define the distribution $G(c)$ over classes as follows:

- For each phonological feature f whose set of possible values is V_f:
  1. Draw $P \sim \text{Bernoulli}(p)$.
  2. If $P = 1$, draw $v \sim \text{Uniform}(V_f)$ and include it in c.
  3. Otherwise, leave the feature unspecified, allowing the class to be abstract.

[4]The particular feature system is treated as a parameter of the model. In the simulation below we used a simplified version of the phonological feature inventory described in Hayes (2011), which only included features that are distinctive in English consonants.

Generating words from templates: Given a choice of phonological template t, we assume that each of the segments that instantiate t has the same probability of being sampled (cf. the "size principle" of Tenenbaum and Griffiths (2001)). Consider again the class c_0 = [+continuant, labial]. Under the assumption that the model's segment inventory is the English one, there are only two segments that are labial continuants: [v] and [f]. The probability of each one of them being generated from c_0 will be $P(s|c_0) = 1/2$.

Prior over template sets: The sounds of the language can be generated from a variety of templates at varying levels of abstraction. We therefore extend the model to be a mixture of template distributions of the type described above. The number of templates is inferred from the data using a Dirichlet process mixture model (Antoniak, 1974).

This prior can be constructed as a process. Suppose that s_i is an ordering of the input sounds, and that we know which templates generated the first $n-1$ sounds $s_1, \ldots, s_{n-1}$. If K is the number of templates that have been posited so far and $n_1, \ldots, n_K$ indicate the number of sounds that have been drawn from each template, then the probability distribution over the template z_n that the sound s_n will be drawn from is given by

$$P(z_n = k|z_{1:n-1}) \propto \begin{cases} n_k & \text{if } k \leq K \\ \alpha & \text{otherwise} \end{cases} \quad (1)$$

Since the probability that an existing template generated s_n is proportional to the number of segments currently assigned to that template, this prior encourages partitions in which a few templates explain most of the sounds (the "rich get richer" property), which amounts to a parsimony bias. Higher values of α can make this bias weaker.

Modeling phoneme spreading: To simulate the generalization made by participants in the IDENTITY experiment, templates must be able to state that two phonemes need to be identical. This is analogous to mechanisms of "spreading" widely assumed in phonology (Colavin et al., 2010; Goldsmith, 1976; McCarthy, 1986). For our simulations below, we simplify by only considering identity constraints between the initial and medial consonants in exposure and test forms. We sample a template over these positions as follows:

1. Draw a class $c_1 \sim G$, where G is the distribution over phonological classes defined above.
2. Draw $Q \sim \text{Bernoulli}(q)$.
3. If $Q = 1$, return an identity template, i.e., $\langle s, s \rangle$ such that $s \in c_1$.
4. Otherwise, draw $c_2 \sim G$ and return the template $\langle s_1, s_2 \rangle$ such that $s_1 \in c_1$ and $s_2 \in c_2$.

Inference: We perform inference to find the posterior over template sets given the exposure datasets used in the human experiments described above. We also infer the hyperparameters using the following prior distributions:

$$\begin{aligned} p &\sim \text{Beta}(1,1) \\ \alpha &\sim \text{Gamma}(2,4) \\ q &\sim \text{Beta}(1,1) \end{aligned} \tag{2}$$

Inference for the Dirichlet process mixture was performed using the Gibbs sampler described in Neal (2000). After each Gibbs sweep, slice sampling (Neal, 2003) was used to obtain a new value for p and q. A new value for α was sampled using the method described by Escobar and West (1995). We ran the sampler for 3000 iterations, discarded the first 100 samples and kept every fifth sample of the remaining samples, for a total of 580 samples from the posterior distribution.

Predicting human data: Participants in the behavioral experiments gave binary judgments ("could the word be part of the language?") rather than probability estimates. To link our model's predictions to participants' binary responses, we sample m template instantiations from the posterior predictive distribution.[5] If the relevant part of the test word appeared in these m samples, the model responds "yes"; this can be understood to be related to a sampling-based view of human inference (Vul et al., 2014). In the simulations below we fix m to be 10.

Human endorsement rates were consistently above 50%, while the model's ratings were often close to 0%. This is likely to be because human ratings were also informed by the unmodeled (fixed) parts of the templates, such as word length or number of vowels. We therefore linearly transform the model's ratings to the range exhibited by human participants: if the untransformed rate is r, the ultimate simulated rate will be $(1 + r)/2$.

[5]Template instantiations only include the modeled (specified) part of the template: an onset consonant in our model of the VOICING language or a consonant pair for the IDENTITY language.

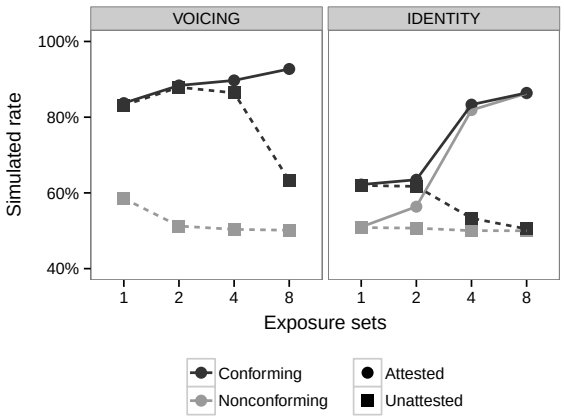

Figure 2: PAIM: Simulated endorsement rates.

4 Simulations

We only modeled those aspects of phonotactic templates that are relevant to the experimental results. For the VOICING experiment, we constrained the template to be ⟨_, V, C, V⟩ (inference is only performed on the first consonant); for the IDENTITY experiment, we constrained it to be ⟨_, V, _, V⟩.

Figure 2 shows the simulated endorsement rates. After a single exposure to each pattern (one exposure set), PAIM behaved in a qualitatively similar way to participants in both experiments: it distinguished CONF from NONCONF words, but did not distinguish ATT from UNATT words.

PAIM was less willing than humans to generalize to CONF-UNATT items after multiple exposure sets: in the IDENTITY experiment the generalization had no effect by the eighth exposure set; in the VOICING experiment its effect was weaker after eight than after four exposures sets. By contrast, human generalization in both languages showed no sign of weakening after multiple exposure sets.

5 Comparison to related models

Hayes and Wilson (2008) propose a Maximum Entropy model of phonotactics (MaxEnt; see also Goldwater and Johnson (2003)). Like PAIM, MaxEnt is based on phonological classes defined as feature matrices. Each class c is assigned a weight w_c. The predicted probability in MaxEnt of a sound s is

$$\hat{p}(s) = \frac{1}{Z} e^{\sum_c w_c I_c(s)} \tag{3}$$

where $Z = \sum_s \hat{p}(s)$ and $I_c(s) = 1$ if $s \in c$ and 0 otherwise.

We simulated endorsement rates from a MaxEnt model for the VOICING language. Following Hayes and Wilson (2008), we used l_2 regularization; that is, if the exposure words were $s_1, \ldots, s_n$, the objective function was

$$\sum_{i=1}^{n} \log \hat{p}(s_i) - \sum_{c} \frac{(w_c - \mu)^2}{2\sigma^2} \qquad (4)$$

Figure 3 shows the simulated endorsement rates for different values of σ (we set $\mu = 0$ in all simulations). For $\sigma = 0.05$, the model showed little learning after a single exposure set. When σ was set to higher values, MaxEnt rapidly preferred attested to unattested items, failing to reproduce the human early generalization pattern. Like PAIM, but unlike humans, generalization to CONF-UNATT items diminished after multiple exposure sets (in particular for $\sigma = 0.5$). A straightforward implementation of MaxEnt is therefore incapable of simulating the human results; better results could potentially be achieved with a regularization method that encouraged sparsity (Goodman, 2004; Johnson et al., 2015).

Another proposed model of phonotactics is the Minimal Generalization Learner, or MGL (Albright, 2009); Linzen and Gallagher (2014) showed that MGL can simulate relevant human behavioral data in some circumstances. In contrast with PAIM and MaxEnt, which converge to the empirical distribution given sufficient data, MGL reserves a fixed amount of probability mass to unseen events. It would therefore able to simulate a sustained generalization pattern.

Our prior over phonological classes bears some resemblance to the Rational Rules model of visual categorization (Goodman et al., 2008). In that model, classes are generated from a probabilistic context free grammar (PCFG); highly specified rules are therefore implicitly less probable, as in our model. Relatedly, Hayes and Wilson (2008) use a greedy feature selection procedure that starts from simpler phonological classes and gradually adds more complex ones; this procedure also encodes an implicit bias in favor of simple classes. Finally, our implementation of a parsimony bias using a Dirichlet process is related to similar biases incorporated into other models of language learning (Frank and Tenenbaum, 2011; Johnson et al., 2007; O'Donnell, 2015).

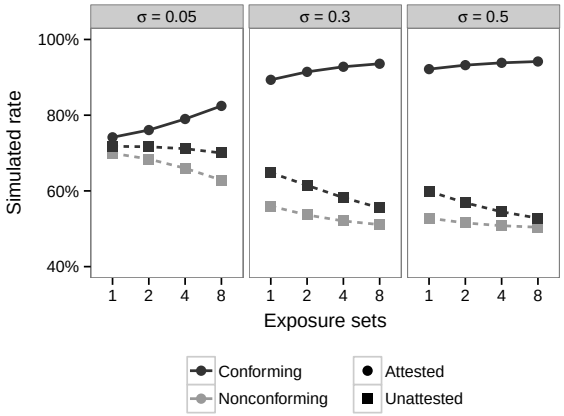

Figure 3: Maximum entropy model: Simulated endorsement rates for the VOICING language, with different values of the regularization parameter σ.

6 Discussion

We have presented a probabilistic model of phonotactic generalization that captures the pattern of rapid, abstract generalization characteristic of human learners. The model's performance is driven by two crucial assumptions. First, it allows hypotheses that make reference to abstract, broad classes of phones from the beginning of the learning process. Second, it prefers to learn compact or parsimonious explanations of the input corpus, using a small number of phonotactic templates. This second property is enforced by our use of the Dirichlet process as a prior over template sets.

These two properties interact. When the model has seen only a few data items, the availability of abstract generalizations allows it to explain all items using a single template, and the prior bias towards parsimony drives it to do so. As the number of data items increases, repeated instances of specific phonemes no longer seem like accidental observations from a more general template, but rather like significant templates in their own right; the model begins to capture such item-specificity.

The model stopped generalizing earlier than humans did; we intend to explore ways to explain this discrepancy. Additional human data would need to be collected to determine whether humans keep generalizing indefinitely, or eventually converge on the attested sounds. Finally, to facilitate inference, we only tested our model on the parts of the word that were relevant to the human data. In future work, we intend to extend the model to learn larger templates that include syllable structure and phonological tiers (Goldsmith, 1976).

References

Adam Albright. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1):9–41.

Charles E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174.

Rebecca S. Colavin, Roger Levy, and Sharon Rose. 2010. Modeling OCP-Place in Amharic with the Maximum Entropy phonotactic learner. In *Proceedings of the 46th meeting of the Chicago Linguistics Society*.

Michael D. Escobar and Mike West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Michael C. Frank and Joshua B. Tenenbaum. 2011. Three ideal observer models for rule learning in simple languages. *Cognition*, 120(3):360–371.

John A. Goldsmith. 1976. *Autosegmental Phonology*. Ph.D. thesis, MIT.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, pages 111–120.

Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. 2008. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154.

Joshua Goodman. 2004. Exponential priors for maximum entropy models. In *HLT-NAACL*, pages 305–312.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Bruce Hayes. 2011. *Introductory phonology*. Wiley-Blackwell, Malden, MA and Oxford.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19*, Cambridge, MA. MIT Press.

Mark Johnson, Joe Pater, Robert Staubs, and Emmanuel Dupoux. 2015. Sign constraints on feature weights improve a joint model of word segmentation and phonology. In *HLT-NAACL*, pages 303–313.

Tal Linzen and Gillian Gallagher. 2014. The timecourse of generalization in phonotactic learning. In *Proceedings of Phonology 2013*.

Tal Linzen and Gillian Gallagher. 2015. Rapid generalization in phonotactic learning. `http://tallinzen.net/media/papers/linzen_gallagher_2015.pdf`.

John J. McCarthy. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17(2):207–263.

Radford M. Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Radford M. Neal. 2003. Slice sampling. *Annals of Statistics*, 31(3):705–767.

Timothy J. O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. The MIT Press, Cambridge, Massachusetts.

Robert J. Scholes. 1966. *Phonotactic grammaticality*. Mouton, The Hague.

Joshua B. Tenenbaum and Thomas L. Griffiths. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640.

Edward Vul, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637.