The Timecourse of Generalization in Phonotactic Learning[*]

Tal Linzen and Gillian Gallagher
New York University

1 Introduction

Native speakers are sensitive to the phonotactic properties of their language, distinguishing nonce words with attested phonotactic structures from nonce words with unattested structures. Sensitivity to phonotactics has been demonstrated for restrictions on sequences of segments, such as syllable onsets in English (Scholes 1966, Berent et al. 2007, Albright 2009, Daland et al. 2011), as well as for non-local cooccurrence restrictions, as in Semitic and Quechua (Berent and Shimron 1997, Frisch and Zawaydeh 2001, Rose and King 2007, Gallagher 2013). Among attested structures, high frequency structures are further distinguished from low frequency ones. Such phonotactic frequency effects have been demonstrated in implicit as well as explicit tasks, in both adults and infants (Jusczyk et al. 1994, Vitevitch et al. 1997, Bailey and Hahn 2001).

In addition to showing sensitivity to attestation and frequency of phonotactic structures, speakers also distinguish among different unattested structures, which all have a frequency of 0. English speakers, for example, prefer unattested onset clusters that share many features with attested onsets to those that have fewer or no features in common with attested onsets. For example, [sɹ] and [vk] are both unattested syllable onsets, but nonce words with the onset [sɹ], which shares most of its features with attested onsets such as [sl] and [ʃɹ], are rated as more acceptable than nonce words with the onset [vk], which is less similar to attested onsets (Scholes 1966, Daland et al. 2011). Likewise, Hebrew speakers judge a nonce root that starts with two identical, non-native consonants, such as [θ-θ], as less acceptable than an analogous root with non-identical consonants (e.g. [θ-k]), even though both sequences have a frequency of 0 in Hebrew (Berent et al. 2002). These results suggest that speakers generalize over attested structures and apply these generalizations to make distinctions among unattested ones. In order to capture speakers' ability to generalize phonotactic knowledge to unattested sequences, models of phonotactic learning have recently begun to go beyond the simple probability of individual structures and represent generalizations over classes of structures (Hayes and Wilson 2008, Albright 2009, Adriaans and Kager 2010, Berent et al. 2012).

While adult speakers clearly learn and use these generalizations, the process by which they are acquired is not well understood. Existing models of phonotactic learning can be classified into two types with respect to the assumptions they make about the order of acquisition of segment-specific knowledge and broader generalizations. In one class of models, which we will refer to as *specific-before-general models*, learners first acquire knowledge about specific segment sequences. Once the learners have noticed the commonalities among the sequences, they form a generalization that (potentially) encompasses some unattested structures in addition to the attested structures that gave rise to the generalization. For example, when acquiring the phonotactics of English, learners must first learn that English syllables can start with [b] and that they can start with [g] before they can make the generalization that English syllables can start with a voiced stop. This generalization can in turn apply to [d], regardless of whether the learner has seen any instances of syllable initial [d]. The Minimal Generalization Learner (Albright and Hayes 2003, Albright 2009) and StaGe (Adriaans and Kager 2010) both fall into this category. In the second class of models, the *simultaneous models*, generalization is not assumed to be a temporal process. The existence of a [b]-initial syllable in the input can support both a segment-specific statement, namely that syllables can start with [b],

[*] For feedback and discussion that led to many improvements to this work we would like to thank Frans Adriaans, Todd Gureckis, Sharon Peperkamp, the NYU Phonetics and Experimental Phonology lab, as well as the audience at and four anonymous reviewers for the Phonology 2013 conference.

Proceedings of *Phonology 2013*
Completed December 14, 2013

and the wider generalization that syllables can start with a voiced stop. This class of models includes Maximum Entropy models (Hayes and Wilson 2008, Pater and Moreton 2012).

This paper reports on an artificial grammar learning experiment designed to provide empirical evidence for choosing between these two classes of models. Our participants learned an artificial language that had both segment-specific phonotactic patterns and a general pattern that applied to a class of segments. All participants were exposed to the same patterns, but the amount of exposure to the language was varied across participants. To anticipate our results, all groups of participants learned the generalization and extended it to unseen items, even with minimal exposure. By contrast, participants were only able to identify individual segment patterns with more exposure. This pattern casts doubt on the specific-before-general assumption: instead of tracking individual structures and then generalizing over those structures, our participants entertained broad generalizations at least as early as specific ones.

2 Experiment 1: Learning an identity generalization

We constructed a language in which all of the words were of the form $C_1V_1C_2V_2$, e.g., *kesa*. This language had two types of phonotactic patterns. Some of the patterns, which we refer to as *arbitrary* patterns, applied to specific segment sequences. For instance, one such pattern would be "$C_1 = k$, $C_2 = s$". In addition to these arbitrary patterns, the language included a general pattern that applied to multiple segment sequences, specifically "$C_1 = C_2$". We presented this language to five groups of participants, who received varying degrees of exposure to the language.

To test participants' learning of the specific and general patterns in the training data, we presented them with new words that had either C_1-C_2 pairs that were seen in training or new C_1-C_2 pairs, and asked them to judge whether the testing words could belong to the language they had learned. Half of the new C_1-C_2 pairs in testing had identical consonants and half had non-identical, resulting in a crossed design that allowed us to test for the independent contribution of the broad and specific generalizations. If participants learned the broad generalization that identical consonant pairs are particularly common in the language, they should prefer items with identical consonant pairs to items with non-identical consonants. If participants learned the segment-specific C_1-C_2 combinations, they should prefer items with attested C_1-C_2 pairs to items with unattested ones. Finally, if participants learned both the broad and specific generalizations, they should prefer words with attested pairs to unattested pairs, and, at least among unattested pairs, should prefer identical to non-identical pairs.

With regards to the main question of interest, the effect of amount of exposure on learning, there are three possible patterns of results. First, participants may show evidence of learning the arbitrary patterns with a small amount of exposure and only start showing evidence of learning the broad pattern with more exposure. This outcome would be straightforwardly consistent with specific-before-general models. Second, participants may show evidence of learning the broad pattern *before* they show evidence of learning the arbitrary patterns, which would favor the simultaneous models over the specific-before-general ones. Finally, participants may start showing evidence of having learned both types of patterns at the same time. This result would be compatible with both types of models.

2.1 *Methods* **2.1.1** *Materials* All words in the experiment were of the form $C_1V_1C_2V_2$, e.g., *kesa*. The training words had one of 8 different C_1-C_2 pairs, 4 of which were identical and 4 of which were not (see Table 1). The testing words had one of 16 C_1-C_2 pairs, 8 containing the consonant pairs heard in training, and 8 containing new consonant pairs. C_1 and C_2 in the novel consonant pairs had always been heard in training, in both initial and medial position, but the C_1-C_2 *combination* had not been heard. A total of 12 unique words were constructed for each consonant pair, by crossing the pair with all non-identical combinations of [a e i u] in V_1-V_2; e.g., for [p-p], the words constructed were *pipa, pipe, pupa* and so on.

The words were recorded by a female native American English speaker, with stress on the initial syllable. The recordings were made at a sampling rate of 44.1 KHz in a sound-attenuated booth on a Marantz PMD-660 solid state recorder using a head-mounted Audio Technica ATM75 microphone.

2.1.2 *Procedure* The experiment was run using Experigen, a program for running online experiments (Becker & Levine 2010). The instructions were as follows:

> "You are going to listen to some words of a made-up language. You do not need to memorize the words, but you should repeat each word to yourself after listening to it.

After hearing several words from the language, you will be presented with some new words and asked whether they sound like they could belong to the language you were listening to."

TRAINING		TESTING			
		attested in training		unattested in training	
Identical		**identical**		**identical**	
p-p	pipa	p-p	papu	k-k	keku
ʃ-ʃ	ʃuʃe	ʃ-ʃ	ʃiʃe	s-s	sasi
g-g	gagu	g-g	guge	dʒ-dʒ	dʒidʒe
n-n	nuni	n-n	nenu	m-m	mamu
arbitrary		**arbitrary**		**arbitrary**	
k-s	kesa	k-s	kusa	p-n	pina
m-dʒ	midʒe	m-dʒ	midʒa	n-g	nage
dʒ-k	dʒaku	dʒ-k	dʒaki	g-ʃ	gaʃe
s-m	semu	s-m	sami	ʃ-p	ʃipu

Table 1: All consonant pairs used in training and testing for Experiment 1, with randomly selected example tokens.

In each training trial, a "play" button appeared in the browser window. When the participant clicked "play", a "continue" button appeared. The participant then clicked "continue" to move on to the next trial. Once the training period was completed, the following instructions screen was displayed: "Now you will be presented with new words and you must decide if they sound like they could belong to the language that you have been listening to." During testing, participants again pressed "play" to listen to the word. After they listened to the word, the question "Does this sound like it could be a word of the language you were listening to?" appeared, along with "yes" and "no" buttons.

In training, participants were assigned to one of five exposure groups. Depending on their group, participants were presented auditorily with 1, 2, 4, 8 or 16 words with each of the 8 training consonant pairs (in total 8, 16, 32, 64 and 128 training words, respectively). The list of words was constructed in blocks, such that each consecutive block of 8 words had exactly one word with each consonant pair; however, participants did not receive any indication of the structure of the lists. The order of consonant pairs within each block and the specific vowel pattern used with each consonant pair was varied across participants.

In testing, participants of all exposure groups were presented with 16 testing words, one word with each consonant pair (see Table 1). The specific word representing each consonant pair again varied across participants, and so did the order of presentation of the words, with the constraint that each block of four consecutive words had exactly one word of each condition (attested identical, unattested identical, attested arbitrary, unattested arbitrary). With the exception of the 16 exposures group, a participant never saw the same word twice; for example, a participant in the 2 exposures group might hear *pipa* and *papu* in training, and *pepi* in testing. The task could therefore not be performed by memorizing the individual words seen in training: participants could only distinguish the legal consonant pairs from the illegal ones by extracting the phonotactic patterns over the consonants. Since there were only 12 words with each consonant pair, training tokens needed to be repeated in the 16 exposures condition.

2.1.3 *Participants* Participants were recruited via Amazon Mechanical Turk (www.mturk.com), and were paid $0.65 for completing the experiment. The experiment took between 2 and 10 minutes, depending on the number of training trials. Participants were told that they needed to be native speakers of English to complete the experiment, and were asked in a short demographic survey at the end of the experiment what their native language was. Participants were limited to those with IP addresses within the United States.

2.1.4 *Statistical analysis* Logistic mixed effects models (LMEM) were fit to the participants' responses ("yes" or "no") using the *lme4* package in R (Bates et al. 2013). We fit a separate model for each exposure group. Contrast coding was used for both consonant pair type (with identical consonants coded as 1 and non-identical as -1) and attestedness (attested: 1, unattested: -1). The models had a maximal random effect structure: for subjects, random intercepts and random slopes for attestedness, identity and their interaction; for consonant pairs, only intercepts. We additionally fit a larger model that included exposure group as a factor. In this larger model, the random effect structure for subjects was identical; for consonant pairs, we added an exposure group random slope. We calculated *p*-values in two ways: using the Wald statistic (i.e. assuming that the distribution of regression coefficients under the null hypothesis is normal), and using the chi-square approximation to likelihood ratio tests in a stepwise regression (similar to a Type I ANOVA), with the predictors entered in the following order: attestedness, consonant pair type, interaction (the order in which the variables are entered should have a negligible effect because our design was orthogonal). The two types of *p*-values converged on the same qualitative results; in the text we report the *p*-values derived from the Wald test (except where noted).

2.2 *Results* Figure 1 shows the mean endorsement rates (proportion of the times that participants judged that the test word belonged to the language they had learned), for the four conditions in each of the five participant groups. We first present the results of LMEMs fitted to each of the five exposure conditions separately (see Figure 2) and then discuss the large LMEM fit to the entire data.

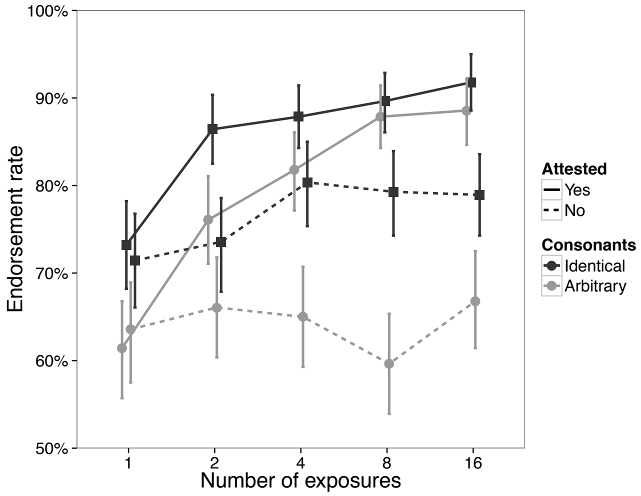

Figure 1: Proportion of "yes" responses in Experiment 1 to the four types of testing items (identical attested = solid dark line, identical unattested = dashed dark line, arbitrary attested = solid light line, arbitrary unattested = dashed light line) in five conditions (1, 2, 4, 8 or 16 exposures to each of the 8 C_1-C_2 pairs in training). Error bars represent bootstrapped 95% confidence intervals.

Participants in the 1 exposure group did not show evidence of distinguishing between attested and unattested items (β = -0.02, p = 0.79), preferred test items with identical consonants to items with nonidentical consonants, indicating that participants learned the identity generalization. The identity preference didn't differ significantly between the attested and unattested pairs (β = 0.07, p = 0.32 for the interaction term between attestedness and identity).

As the number of exposures to the items increased, participants started to favor attested items. Even the participants who saw only 2 words with each consonant pair showed a reliable preference for the attested items (β = 0.39, $p < 0.001$). In other words, with two or more exposures to each item, participants showed evidence of keeping track of the individual C_1-C_2 pairs presented in training. The effect of identity on endorsement rate persisted in the 2 exposure condition, and again did not differ significantly between attested and unattested items, though there was a numerically larger effect of identity for attested items (interaction: $p_z = 0.1$, $p_{\chi^2} = 0.18$). Since the effects of attestedness and identity had similar magnitude, unattested identical items and attested arbitrary items were rated similarly. The pattern remained statistically similar in the 4 exposure condition, though the (still non-significant) interaction switches direction, such that identity starts to play a larger role in unattested items than in attested items.

The preference for attested consonant pairs increased steadily, becoming twice as strong in the 16 exposures condition as in the 2 exposures condition (16 exposures: β = 0.77, $p < 0.001$). Also by the 16 exposures condition, identity stopped having an effect on attested items: once participants have learned the attested consonant pairs, they no longer rely on identity to judge their well-formedness. The preference for identical pairs persists in the unattested consonant pairs, driving a main effect of identity; in the 8 exposures condition, this main effect is offset by a significant interaction, which confirms that it is specific to the unattested pairs (β = -0.27, *p* = 0.006). However, the unusually large effect of identity in the 8 exposures condition appears to be driven by a potentially spurious reduction in the endorsement rate of the unattested arbitrary items; in the 16 exposures condition, the magnitude of the identity effect returns to its level from the 4 exposure condition, and the interaction is no longer significant (though still trends in the same direction).

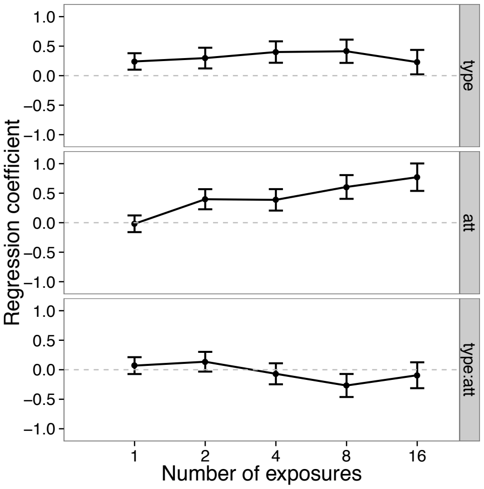

Figure 2: Logistic MEM coefficients for the models fit to each exposure group separately. Error bars indicate two standard errors of the estimates.

We additionally fitted LMEMs to the entire data set, with the number of exposures as a predictor in the regression model. We fitted two models, coding the number of exposures in two different ways: as a categorical variable with 5 levels, and as a linear trend, using a logarithmic transformation (i.e., the exposure groups 1, 2, 4, 8 and 16 were transformed to 1, 2, 3, 4 and 5 respectively). All three main effects (consonant pair type, attestedness and number of exposures) were highly significant in both models. The interaction between attestedness and number of exposures was also significant, reflecting the participants' consistent improvement in discriminating between attested and unattested items ($p < 0.001$).

The overall interaction between attestedness and type was not significant in either model. The interaction between number of exposures and type was not significant either: the identity generalization is learned very quickly, and there is no improvement with additional exposure. However, the interpretation of this term is complicated by a marginally significant three-way interaction between attestedness, type and number of exposures (*p* = 0.08 in the categorical model, *p* = 0.05 in the linear trend model). This pattern reflects the reversal of the trend of the type-by-attestedness interaction in the group-by-group models: it starts out positive, such that identical attested items are learned slightly better than non-identical attested items (1 and 2 exposures), and gradually becomes negative (8 and 16 exposures), such that identity stops conferring an advantage for attested items and comes into play only when the participants judge unattested items.

2.3 *Discussion* Participants showed evidence of learning the broad generalization over identical consonant pairs before they showed evidence of learning the individual attested C_1-C_2 pairs. After a single exposure to each of 8 possible consonant pairs, 4 of which were pairs of identical consonants, participants showed a preference for novel words with identical consonants. This preference held regardless of whether or not this pair of identical consonants was presented in training. Participants did not start showing evidence of learning individual consonant pairs until they had 2 exposures to each type. This pattern of results suggests that general patterns can be learned before individual instances of these patterns, contra the specific-before-general hypothesis.

Average endorsement rates were typically above 60%, indicating that participants did not strongly object to any of the test items. This is not particularly surprising given that all test items followed the template $C_1V_1C_2V_2$, which was shared by all training items, and were made up of the same consonants and vowels as the training items. Thus, all of the testing items received some support from the training data.

Items with novel syllable structures (e.g., [kes]) or novel segments (e.g., [lumi]) would likely be endorsed at a lower rate.

3 Experiment 2: Ruling out a pre-existing bias

We interpreted our participants' preference for identical items after one exposure in Experiment 1 as reflecting the learning of a generalization. Before being confident in this interpretation, however, we must rule out the possibility that these results are due to prior bias in favor of words with identical consonants. To confirm that the preference for identical consonant pairs after one exposure in Experiment 1 was due to training, we ran Experiment 2 as a control. In Experiment 2, participants were exposed to 8 arbitrary (i.e., non-identical) C_1-C_2 pairs, and then tested on the same unattested items as in Experiment 1 (including the identical ones). If participants still show a preference for identical over non-identical items, despite not having seen any identical items in training, this will be evidence that the preference is due to some prior bias in favor of identical items. If, however, participants show no identity preference, this will support the interpretation of the identity preference in Experiment 1 as a result of learning.

3.1 *Methods* **3.1.1** *Materials* All words had the form $C_1V_1C_2V_2$, as in Experiment 1. As in the 1 exposure condition of Experiment 1, there were 8 training words and 16 test words. All training words had non-identical consonants (see Table 2). Vowel patterns were chosen at random, with no vowel pattern repeated across training and testing words. As in Experiment 1, half of the test words were attested in training and half weren't. All of the attested words in testing had non-identical consonants, resulting in a necessary mismatch with Experiment 1. The unattested items in testing had the same consonant pairs as in Experiment 1, half identical and half non-identical (four of each).

The support that identical and non-identical test items could receive from accidental patterns in the training set was matched as follows. Each of the eight segments appeared once in initial position and once in medial position. The identical and non-identical unattested test items therefore received equal support from the positional frequency of the individual segments, as in Experiment 1. In addition, the two types of unattested test items were matched for the amount of natural class based support they received from consonant cooccurences in the training items (voicing, place of articulation and manner of articulation). For example, the test item s-s receives support from two voiceless-voiceless pairs (p-s and k-p), and there are no fricative-fricative pairs or alveolar-alveolar pairs in the training data, so its total natural class-based cooccurence support score is 2. It is matched with g-ʃ, which also receives natural-class based support from 2 attested pairs, the single stop-fricative pair p-s and the single voiced-voiceless pair g-k; there are no velar-palatal pairs in the training data.

TRAINING		TESTING			
		attested in training		unattested in training	
ʃ-dʒ	ʃadʒi	ʃ-dʒ	ʃidʒe	**identical**	
m-n	mena	m-n	muni	k-k	keku
s-g	sagu	s-g	sage	s-s	sasi
p-s	pesi	p-s	pisu	dʒ-dʒ	dʒidʒe
g-k	giku	g-k	guka	m-m	mamu
k-p	kape	k-p	kepi	**arbitrary**	
n-ʃ	nuʃa	n-ʃ	nuʃa	p-n	pina
dʒ-m	dʒemu	dʒ-m	dʒamu	n-g	nage
				g-ʃ	gaʃe
				ʃ -p	ʃipu

Table 2: All consonant pairs used in training and testing for Experiment 2, with randomly selected example tokens.

3.1.2 *Participants* Participants were recruited in the same way as in Experiment 1. There were 70 participants.

3.1.3 *Procedure* The procedure was identical to Experiment 1.

3.1.4 *Statistical analysis* A logistic mixed-effects model was fit to the results, with a three-level factor of consonant type (attested, unattested identical, unattested arbitrary) as a fixed effect and a by-subject random effect, as well as by-subject and by-item random intercepts. All p-values are derived from the Wald test.

3.2 *Results* The results of Experiment 2 are shown in Figure 3. Contrary to the predictions of the bias hypothesis, participants did not show a preference for identical unattested items; if anything, there was a slight preference for non-identical unattested items. Incidentally, there was a striking difference between the attested (non-identical) items and unattested items of either type, however: unlike in the 1-exposure condition of Experiment 1, participants were much more likely to endorse attested than unattested items. The effect of type was highly significant ($p < 0.001$). Planned comparisons showed that this effect was entirely due to the difference between attested and unattested items: the difference between the two types of unattested items was far from being significant ($p = 0.55$).

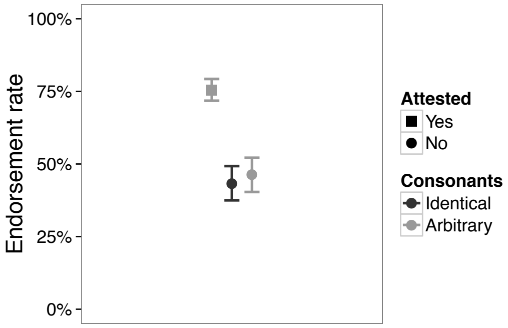

Figure 3: Proportion of "yes" responses in Experiment 2 to the three types of testing items (attested = dark triangle, arbitrary unattested = dark circle, identical unattested = light circle), with one exposure to each consonant pair in training. Error bars represent bootstrapped 95% confidence intervals.

3.3 *Discussion* When participants were not exposed to identical consonant pairs in Experiment 2, they did not show any preference for novel items with identical consonants. The results therefore support interpreting the preference for identical items after one exposure in Experiment 1 as being due to learning during the training portion of the experiment. Thus, the interpretation of the main result of Experiment 1 remains unchanged: participants show evidence of learning the broad generalization about identical consonant pairs before learning narrow generalizations about the specific attested consonant pairs.

The results of Experiment 2 reveal an additional interesting effect. Participants in Experiment 2 showed a strong preference for attested over unattested consonant pairs after just one exposure, unlike participants in Experiment 1. While we cannot make firm claims about the source of this difference, one possibility is that the presence of a broad generalization interferes with the learning of narrower generalizations. In Experiment 1, the presence of the identity generalization prevented learners from attending sufficiently to the narrower generalizations with small amounts of training, while in Experiment 2 learners were free to focus on the specific, attested C_1-C_2 pairs.

At first blush, the lack of a preference for identical items in Experiment 2 compared to Experiment 1 could still be consistent with a pre-existing bias to give "yes" responses to identical items: the absence of identical consonant pairs in the training data could have been taken as evidence for the generalization that pairs of identical consonants are *underattested*, offsetting a pre-existing bias in favor of identical consonants. However, this alternative explanation for the results of Experiment 2 becomes less plausible if we consider the radically different amount of support for the generalization that the training data provide in each of the experiments. With an inventory of 8 consonants, a sample of 8 words with all non-identical pairs is not a particularly surprising one: 56 out of the possible 64 consonant pairs are non-identical. The expected number of non-identical pairs in a sample of 8 is therefore 7, and an observed sample of 8 non-identical items yields an observed-over-expected ratio (O/E) of 8/7. In Experiment 1, on the other hand, the participants received 4 identical pairs instead of the expected 1, for an O/E of 4/1. In other words, the evidence for the overattestation of identical pairs in Experiment 1 is much stronger than the evidence for their underattestation in Experiment 2. It is therefore implausible that the preference for identical items after one exposure in Experiment 1 was due to bias, and at the same time that the lack of preference for identical items in Experiment 2 was due to learning that offset the bias.

4 General discussion

We found that participants show evidence of learning a broad generalization (pairs of identical consonants are attested) before learning more narrow generalizations (which individual segmental pairs are present in the training data), calling into question the specific-before-general assumption. Our results do not present direct evidence that broad generalizations are explicitly favored over more narrow generalizations in learning. In any sample size, the amount of evidence for the broad and narrow generalizations differ. Broad generalizations are typically supported by a greater number of types than narrow generalizations, so it is expected that learners would learn them faster, all other things being equal. Specifically, after a single exposure to each of the 8 consonant pairs, the learner has seen 4 tokens that support the broad generalization about identical consonants, and only one training token that supports the narrow [k-s] generalization. Even assuming that the learner is not biased towards either general or specific patterns, the identity pattern would be learned first, as long as both general and specific hypotheses are considered simultaneously.

4.1 *Relation to previous empirical work* Our results help explain some divergent findings from the artificial grammar learning literature. In a recent study with 6-month-old infants (Cristia and Peperkamp 2012), participants were exposed to $C_1V_1C_2V_2$ words with three different consonants in C_1 (18 words with each onset). All three consonants had the same voicing (either all voiced or all voiceless). In testing, the infants listened to words with 1) attested onsets with the same voicing as in training, 2) unattested onsets with the same voicing as in training, 3) unattested onsets with different voicing (though no infant heard all three conditions). The infants looked longer to trials of condition 2 than condition 3, indicating that they learned the voicing generalization; however, they did not distinguish condition 1 from condition 2, suggesting that they did not store the individual consonants. This contrasts with previous adult studies in which participants did not generalize at all beyond the individual segments they saw in training (Peperkamp, Skoruppa and Dupoux 2006; Peperkamp and Dupoux 2007). The authors concluded that infants and adults learn phonotactics differently: infants encode class-wide patterns, whereas adults learn individual segments.

The evidence for a qualitative difference between infants and adults is mixed, however. On the one hand, 9-month-old infants trained on unsegmented speech from an artificial language successfully learned a phonotactic pattern whereby syllable onsets must be voiced the stops /b d g/, but failed to learn a pattern that did not lend itself to a natural-class based grouping, e.g. /b t g/ (Saffran and Thiessen 2003). This result supports Cristia and Pepperkamp's hypothesis that infants are particularly good at learning phonotactics over natural classes. On the other hand, when trained on individual segmented syllables, slightly older infants (10.5 months old) successfully learned an arbitrary phonotactic pattern (Chambers, Onishi and Fisher 2011).

Further undermining the dichotomy between adults and infants, adults do show generalization to novel segments in some experiments, though they typically prefer segments seen in training to unseen but generalization-conforming segments (Finely and Badecker 2009, Cristia et al. 2013, Gallagher 2013). In one study (Cristia et al. 2013), participants were exposed to words of the form $C_1V_1C_2V_2$, where C_1 was one of five consonants drawn from a subset of some phonological natural class (for example, /d g v z ʒ/, a subset of the class of voiced obstruents). In the relevant conditions of the test phase, participants were requested to give frequency judgments on novel words in which C_1 was either 1) one of the attested onsets, 2) an unattested onset from the same natural class (e.g., /b/ in the case of voiced obstruents), or 3) an unattested onset that didn't belong to the natural class (e.g., /p/). Participants rated the attested onsets as most frequent, the generalization-conforming unattested onsets as less frequent, and non-conforming unattested onsets as the least frequent. This pattern, which is similar to the pattern we saw in the higher exposure conditions, is the pattern predicted by most extant learning models.

Abstracting away from the specifics of the studies surveyed above, the results of Experiment 1 suggest a potential way to reconcile these conflicting findings. After a single exposure to each consonant pair, adult participants in our experiment learned the generalization that identical consonant pairs are preferrable to non-identical pairs, but failed to learn the individual attested consonant pairs. In other words, when they received a small amount of exposure, adults showed the same performance as the 6-month-old infants in Cristia and Peperkamp's (2012) study. In the 8 and 16 exposure conditions of Experiment 1, on the other hand, our participants' performance mirrored the results of Cristia et al.'s (2013) adult study: attested consonants pairs were rated highest, unattested generalization-conforming items rated lower, and unattested non-conforming items rated lowest.

If each type in the language is experienced with equal frequency, then, general patterns that receive support from multiple types (e.g., "two identical consonants"), are learned faster than patterns that only receive support from a single type (e.g., s-m). This raises the possibility that the difference between infants and adults may not be a qualitative difference in learning strategies or biases, but rather a quantitative difference in the amount of evidence each group receives for the pattern being learned. Adults often receive considerably more exposure to the artificial language than infants, due to the difficulty maintining the infants' attention for extended periods of time. For example, adult participants in Cristia et al. (2013) were exposed to 160 training words, whereas infants in Cristia and Peperkamp (2012) were only exposed to 54 words. Further work is required to establish how much of the differences between infant and adult studies can be explained by the differences in the amount of exposure that they receive.

4.2 *Implications for specific-before-general models* Simultaneous models assume that both general and specific patterns are entertained as potential hypotheses from the outset of the learning process (Hayes and Wilson 2008, Pater and Moreton 2012). These models can straightforwardly account for a scenario in which participants fail to learn a specific pattern but manage to learn a general one, as in the 1- exposure condition of Experiment 1 and in some of the infant studies mentioned above. Specific-before-general models like MGL (Abright 2009) and StaGe (Adriaans and Kager 2010), on the other hand, predict that learners need to first learn specific instances of a pattern before they can form the more general pattern.

In a sense, the result of the 1-exposure condition in Experiment 1 is a null result: it shows that the general pattern was learned better than the specific ones, but does not prove conclusively that participants didn't learn the specific patterns. In principle, these results could be compatible with a specific-before-general model that learned both a specific and a general pattern, but assigns a much higher weight to the general pattern. The effect of the specific patterns could then be too small to detect in our experiment, even though both types of patterns have been learned (i.e., represented). While this is a theoretical possibility, the actual implementations of the MGL and StaGe predict exactly the opposite: when both a specific and a general pattern apply to a test string, the specific pattern typically overrides the general ones. Specifically, MGL assumes that each pattern that applies to the string being evaluated contributes a phonotactic probability estimate. The estimate contributed by each applicable pattern is calculated by dividing the relative frequency of the pattern by the number of types that the pattern could apply to. The overall estimate for a particular string is then obtained by taking the highest estimate assigned by all applicable patterns. StaGe, which is based on Optimality Theory, uses a similar mechanism to determine the rank of the relevant constraint. Again, only the highest ranked constraint has an effect.

To illustrate this procedure for MGL, suppose that the model has been trained on the training set of Experiment 1 and has acquired both the specific patterns (*p-p, k-s,* etc.) and the general pattern $C_1 = C_2$. The probability estimate for the generalization $C_1 = C_2$ will be the relative frequency of identical pairs in the training data divided by the number of possible identical pairs. The relative frequency of identical pairs is 1/2, and there are 8 possible identical pairs; the resulting probability estimate is therefore (1/2)/8 = 1/16. A specific pattern such as *p-p* only applies to one type, which occurs once every 8 words, so its probability estimate would be (1/8)/1 = 1/8, which is higher than the estimate for the more general pattern. The model's well-formedness scores for novel words in Experiment 1 would be as follows:

(1) a. Attested identical (*pipu*): 1/8 (both *p-p* and $C_1 = C_2$ apply; *p-p* has a higher estimate)
b. Attested arbitrary (*kisu*): 1/8 (only *k-s* applies)
c. Unattested identical (*keku*): 1/16 (only $C_1 = C_2$ applies)
d. Unattested arbitrary (*pina*): 0 (no pattern applies[1])

[1] This is a simplification, since *pina* conforms to many generalizations about the segments, natural classes and prosodic structure of the language. It also conforms to the very general cooccurrence pattern C_i-C_j ("any two consonants"), which has a relative frequency of 1 but applies to all 64 possible types, so that it assigns a probability estimate of 1/64.

The estimates in (1) correctly predict the qualitative pattern of results of the 16 exposures group in Experiment 1: attested items are rated highest; unattested generalization-conforming items are rated lower; and unattested arbitrary items are rated lowest. Interestingly, MGL also correctly predicts that the rating of attested pairs does not depend on whether they conform to the generalization or not, a pattern not predicted by Maximum Entropy models.

While MGL successfully predicts the results of the 16 exposures condition, it does not explain participants' behavior in the 1 exposure group: it predicts a preference for specific patterns regardless of the number of tokens supporting each type of pattern. However, the model can be modified to take the number of tokens into account by reintroducing a component proposed in a previous version of the model (Albright and Hayes 2003). Specifically, phonotactic probability can be estimated using the lower bound of the confidence interval (CI) for the relative frequency instead of the relative frequency itself. For example, the relative frequency of a consonant pair is the same whether it was seen 2 out of 16 times or 100 out of 800 times (0.125 in both cases), yet the CI-based estimates would be very different between the two cases: the 95% CI is (0.03, 0.36) in the first case, leading to an estimate of 0.03, and (0.1, 0.15) in the second, leading to an estimate of 0.1. This procedure has a similar effect as regularization in regression models (used, for example, by Hayes and Wilson 2008): probability estimates are pulled towards zero if the data don't provide strong enough support for them.

We ran simulations of the modified version of MGL in which probability estimates contributed by a generalization are adjusted depending on the number of tokens the generalization is based on. Adjusting the model's probability estimates based on the amount of evidence may in some cases allow the general identity pattern to be weighted higher than the specific patterns, despite the fact that the identity pattern has a lower unadjusted estimate than the specific ones. For instance, the unadjusted estimate for a specific pattern may be 0.125, but could be adjusted to 0.01 if it is based on just one token. By contrast, the estimate for the general pattern, which has a lower unadjusted value (0.0625) but is based on more tokens, would be adjusted less dramatically, say to 0.05. Figure 4 shows the result of simulations of an MGL phonotactic learner exposed to the training data from Experiment 1, with different CI sizes. When the CI size α is small (e.g., α = 25%), the estimate is close to the relative frequency even after one exposure. When it is large (e.g., α = 95%), the estimates are pulled heavily towards zero, and a greater amount of types is required for the relative frequency pattern to have an effect. When α is set to 75%, the simulation results match the qualitative pattern of results from Experiment 1.

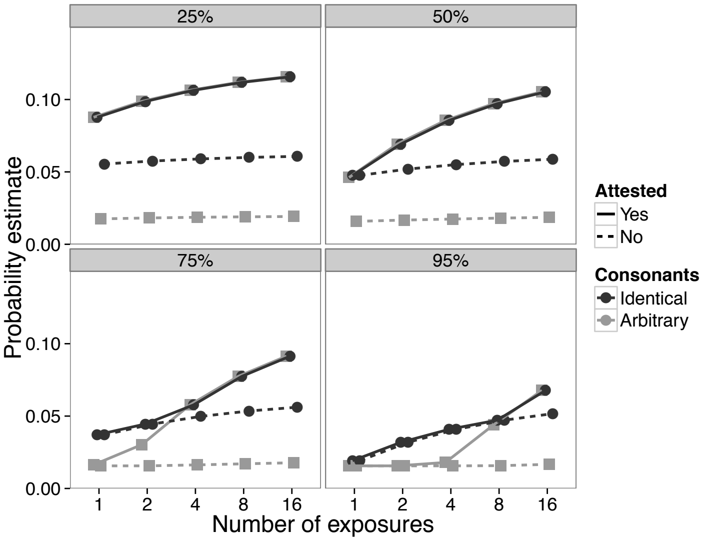

Figure 4: Phonotactic probability estimates from a modified version of the MGL phonotactic learner (Albright 2009) in which a probability estimate is replaced by the lower bound of a confidence interval around the estimate, for four sizes of the confidence interval (25%, 50%, 75% and 95%).

To summarize, it is possible to modify specific-before-general models to simulate the results of Experiment 1, based on the assumption that participants do learn specific patterns immediately and use them to form more general patterns, but don't show evidence of having learned the specific patterns until they've seen multiple tokens instantiating them. In other words, we need to attribute knowledge to the learners before they show evidence of applying it, which may be theoretically undesirable. Furthermore, the fact that different values of α can lead the models to predict very different patterns of results raises the concern that this modified model may not be restrictive enough. Albright and Hayes (2003) mention that their model achieved the best fit to the English past tense data with α = 55%. In our data set, however, α needs to be closer to 75%. It would be preferable to have a principled way to predict what α should be for a given data set.

An issue worth mentioning is the way in which identity generalizations are represented in the model. MGL as described in Albright (2009) is a model of natural class based generalization. After the model has learned the patterns *p-l* and *b-l* (or, in simplified phonological feature notation, [lab, –voice]-*l* and [lab, +voice]-*l*) it notices the natural classes that these patterns have in common, abstracts over the differences between them and induces the pattern [lab]-*l*. This generalization procedure does not straightforwardly extend to identity generalizations. For instance, *p-p* and *b-b* would be represented as [lab, –voice][lab, –voice] and [lab, +voice][lab, +voice] respectively; the only generalization that MGL would extract from these two pairs of feature bundles is [lab][lab], a pattern that doesn't encode the identity between the two consonants, and therefore incorrectly applies to *p-b* as well.

To enable MGL to learn identity patterns, we coded each of the identical input pairs twice, once as a simple pair of feature bundles, and once as a single feature bundle followed by a variable X. The input *p-p*, for example, was coded both as the simple representation [lab, –voice][lab, –voice] and as the variable-based representation [lab, –voice]-X (Colavin et al. 2010, Gallagher 2013). Given this double representation of each identical input, *p-p* and *b-b* can now give rise to two generalizations: [lab][lab], which covers *p-b, b-b, b-p* and *p-p*, and [lab]-X, which only covers *p-p* and *b-b*. Some machinery is needed to deal with identity patterns in Maximum Entropy models as well, though the added complication is concentrated in the set of possible patterns, and does not require recoding the relationship between the consonants as part of the representation of the pair. Specifically, in addition to simple patterns such as "$C_1 = s$, $C_2 = k$", the input needs to be matched against "$C_1 = C_2$" (Berent et al. 2012).

5 Conclusion

The experiments presented in this paper compared the timecourse of acquisition of specific and general phonotactic patterns. We found that the general identity pattern was acquired early on, resulting in a preference for identical over non-identical consonant pairs. The learning of individual consonant pairs proceeded more slowly. Initially, participants didn't reliably distinguish attested from unattested consonant pairs. As the number of exposures increased, participants made a larger and larger distinction between attested and unattested pairs. The earlier acquisition of the general pattern compared to the specific patterns is more straightforwardly explained by simultaneous models, which can acquire both general and specific patterns at the same time, than by specific-before-general models.

Experiment 2 showed that participants do not prefer identical test items if they do not encounter them in the training phase, confirming that participants' preference for identical test items in Experiment 1 was due to the training they received and not to a pre-existing bias in favor of identical consonants. We also found that the individual arbitrary items were learned better in Experiment 2 than in Experiment 1, indicating that the presence of a broad generalization may impair the learning of specific patterns.

Even though the relative frequency of the patterns was identical across the five exposure conditions of Experiment 1, participants showed qualitatively different behavior based on the amount of training data they saw. This sensitivity to the amount of evidence could account for differences between studies that showed learning of specific patterns on the one hand and studies that showed exclusive learning of a general pattern (particularly in infants) on the other hand. We also showed that given certain assumptions specific-before-general models could be modified to accommodate the pattern we found, though simultaneous models offer a more straightforward account of the results. In general, our results point to the potential of timecourse data as a tool to constrain models of phonotactic learning.

References

Adriaans, Frans and René Kager (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language, 62*, 311–331.

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

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

Becker, Michael and Jonathan Levine (2010). Experigen: an online experiment platform. Available at https://github.com/tlozoot/experigen.

Bates, Douglas, Martin Maechler, and Ben Bolker (2013). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-2.

Berent, Iris and Joseph Shimron (1997). The Representation of Hebrew Words: Evidence from the Obligatory Contour Principle. *Cognition, 64*, 39-72.

Berent, Iris, Gary Marcus, Joseph Shimron, and Adamantios Gafos (2002). The scope of linguistic generalizations: Evidence from Hebrew word formation. *Cognition, 83*, 113–139.

Berent, Iris, Colin Wilson, Gary Marcus and Doug Bemis (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry, 43*, 97–119.

Chambers, Kyle E., Kristine H. Onishi, and Cynthia Fisher. 2011. Representations for phonotactic learning in infancy. *Language Learning and Development, 7*, 287–308.

Colavin, Rebecca S., Roger Levy, and Sharon Rose (2010). Modeling OCP-Place in Amharic with the Maximum Entropy phonotactic learner. *Chicago Linguistic Society, 46*.

Cristia, Alejandrina and Sharon Peperkamp (2012). Generalizing without encoding specifics: Infants infer phonotactic patterns on sound classes. *Proceedings of the 36th Annual Boston University Conference on Language Development (BUCLD 36),* 126–138.

Cristia, Alejandrina, Jeff Mielke, Robert Daland, and Sharon Peperkamp (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, *4*, 259–285.

Finley, Sara, and William Badecker (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, *61*, 423–437.

Gallagher, Gillian (2013). Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, *30*, 253–295.

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

Jusczyk, Peter W., Paul A. Luce, and Jan Charles-Luce (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.

Pater, Joe and Elliott Moreton (2012). Structurally biased phonology: complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad*, *3*, 1–41.

Peperkamp, Sharon and Emanuel Dupoux (2007). Learning the mapping from surface to underlying representations in an artificial language. In *Laboratory Phonology IX*, ed. Jennifer Cole and Jose-Ignacio Hualde, 315–338. Berlin: Mouton de Gruyter.

Peperkamp, Sharon, Katrin Skoruppa, and Emanuel Dupoux (2006). The role of phonetic naturalness in phonological rule acquisition. In *Proceedings of the 30th Annual Boston University Conference on Language Development*, ed. D. Bamman, T. Magnitskaia, and C. Zaller, 464–475. Cascadilla Press.

Saffran, Jennifer, and Eric D. Thiessen (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*, 484–494.

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

Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce, and David Kemmerer (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, *40*, 47–62.