Rapid generalization in phonotactic learning

Tal Linzen[1,2] and Gillian Gallagher[3]

[1]LSCP & IJN, CNRS, ENS, PSL Research University
[2]Department of Cognitive Science, Johns Hopkins University
[3]Department of Linguistics, New York University

June 2, 2017

Abstract

Speakers judge novel strings to be better potential words of their language if those strings consist of sound sequences that are attested in the language. These intuitions are often generalized to new sequences that share some properties with attested ones: participants exposed to an artificial language where all words start with the voiced stops [b] and [d] will prefer words that start with other voiced stops (e.g., [g]) to words that start with vowels or nasals. The current study tracks the evolution of generalization across sounds during the early stages of artificial language learning. In Experiments 1 and 2, participants received varying amounts of exposure to an artificial language. Learners rapidly generalized to new sounds: in fact, following short exposure to the language, attested patterns were not distinguished from unattested patterns that were similar in their phonological properties to the attested ones. Following additional exposure, participants showed an increasing preference for attested sounds, alongside sustained generalization to unattested ones. Finally, Experiment 3 tested whether participants can rapidly generalize to new sounds based on a single type of sound. We discuss the implications of our results for computational models of phonotactic learning.

1 Introduction[1]

Natural languages typically place restrictions on the ways in which sounds can combine to form words. The consonant [h], for example, can occur in the onset of an English syllable, as in *half* [hæf], but not in its coda: English does not have words like **fah* [fæh] (McMahon, 2002). English speakers do not typically consider this gap to be accidental; they judge words that end with a [h] as unlikely to become words of the language. The set of all such restrictions is referred to as the phonotactics of the language. The distinction between

[1]All code and data necessary to reproduce the findings reported in this paper can be found at `http://github.com/TalLinzen/rapid_phonotactic_generalization`.

Most of this work was done when the first author was at New York University. Experiments 2a and 2b were published in the Proceedings of Phonology 2013 in a different form (Linzen & Gallagher, 2014). We thank Frans Adriaans, Alex Cristia, Robert Daland, Michael C. Frank, Todd Gureckis and Timothy O'Donnell for discussion, as well as audiences at NYU, Tel Aviv University, MIT, Stanford, LSCP, the Northeast Computational Phonology Circle meetings at Yale and NYU, the 36th Annual Cognitive Science Society Meeting and the 2016 LSA Annual Meeting. Tal Linzen's research was supported by the European Research Council (grant ERC-2011-AdG 295810 BOOTPHON) and the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC).

phonotactically legal and illegal words is reflected in a variety of implicit tasks, in both adults and infants (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; McQueen, 1998).

Sounds with similar articulatory or perceptual properties tend to have similar phonotactic distributions. In German, for example, voiced stops (e.g., [b] or [g]) are not allowed at the end of a syllable: [bal] is a valid German word, but *[lab] is not (Jessen & Ringen, 2002). Speakers often use such class-wide phonotactic patterns to generalize from structures attested in their lexicon to new sounds and sequences. For example, the onsets [sr] and [mb] are both unattested in English, but [sr] is similar to the attested strident-liquid onsets [sl] and [ʃr] while there are no attested sonorant-stop sequences similar to [mb] (Albright, 2009). English speakers judge *srip* to be a better potential word than *mbip* (Scholes, 1966; Daland et al., 2011); this suggests that speakers judge novel structures based on similarity to existing structures.

Phonotactic learning can be productively studied in a controlled setting in artificial language learning experiments. In these experiments, participants are presented with individual words in a miniature artificial language, and are then tested on novel words to determine what knowledge they have extracted from the language. Adult participants have shown evidence of learning the phonotactics of artificial languages in such diverse tasks as acceptability judgments (Richtsmeier, 2011), speech error patterns in production (Gaskell et al., 2014; Warker & Dell, 2006) and familiarity judgments (Cristia, Mielke, Daland, & Peperkamp, 2013). Related findings have been reported in artificial language studies of morphological alternations (Finley & Badecker, 2009; Peperkamp, Le Calvez, Nadal, & Dupoux, 2006; Wilson, 2003). Similar findings have been reported for infants, who are typically tested using the headturn preference paradigm (Chambers, Onishi, Fisher, et al., 2003; Cristià & Seidl, 2008; Saffran & Thiessen, 2003; Seidl & Buckley, 2005; Cristia & Peperkamp, 2012). Most importantly for the present study, multiple artificial language learning experiments have reported generalization beyond attested sounds (Cristia & Peperkamp, 2012; Cristia et al., 2013; Gallagher, 2013; Finley & Badecker, 2009; Marcus, Vijayan, Rao, & Vishton, 1999), though some studies have failed to find this effect (Peperkamp et al., 2006; Peperkamp & Dupoux, 2007).

Phonotactic learning studies with adults usually provide participants with considerable exposure to the language. Likewise, models of phonotactic learning focus on the end stage of the learning process, after a large amount of data from the language has been encountered. Conversely, there is little empirical data and modeling work bearing on the time course of phonotactic learning. How much evidence do learners need to begin generalizing to new sounds? Does generalization to novel sounds that have a particular phonological feature (e.g., voiced stops) require multiple attested sounds that have that feature? In what way does the likelihood of generalization diminish as participants are exposed to more examples of the set of sound sequences that exist in the language? In the rest of this introduction, we describe how the answers to these questions could inform models of phonotactics (Section 1.1) and outline the artificial language learning experiments presented in the rest of the paper, which begin to address these questions (Section 1.2).

1.1 The time course of generalization in models of phonotactics

We first review two prominent views of the time course of generalization to new sounds in probabilistic models of phonotactics (see also Cristia & Peperkamp, 2012; Kapatsinski, 2014). In practice, both families of models are quite flexible, and may be modified to accommodate a wide range of results; while we do not intend this paper as a final adjudication between these two types of models, it would be useful to survey these classes of models as a framework with which to evaluate our experimental results.

One family is based on minimal generalization learning (Albright & Hayes, 2003; Albright, 2009; Adriaans & Kager, 2010). In these models, learners generalize to the smallest phonological class that contains the sounds supporting the generalization. In particular, this means that a single sound does not lead to any generalization. Once learners have noticed the commonalities among multiple sounds they have acquired,

they form a generalization over the smallest phonological class that contains these sounds; this class can include some unattested sounds. For example, when acquiring the phonotactics of English, learners may first learn that both [b] and [g] are valid onsets for English syllables before they can generalize to other voiced stops (e.g., [d]). This generalization will be restricted to the minimal class that contained the attested onsets (i.e. voiced stops), at least until a voiceless stop onset is encountered. This assumption, which we refer to as the *specific-to-general assumption*, is in line with the finding from the artificial language learning literature that infants require three different exposure types to generalize to a novel item (Gerken & Bollt, 2008).

Other models consider representations at multiple levels of generality from the earliest stages of learning, without waiting for multiple sounds to support a particular dimension of generalization (Hayes & Wilson, 2008; Moreton, Pater, & Pertsova, 2015; Linzen & O'Donnell, 2015). In maximum entropy models such as the Hayes and Wilson (2008) model or GMECSS (Moreton et al., 2015), for example, the well-formedness of a sound is derived from a linear combination of the weights associated with each of the phonological classes that the sound belongs to: the well-formedness of a [b] is determined by the weight for [b] and the weight for the class of voiced stops, among the various other classes that [b] belongs to. Learning the phonotactics of the language consists in determining the set of weights that is most consistent with the statistical distribution of sounds in the language. Since both the sound-specific and the class-wide weight contribute to the well-formedness of a [b] token, exposure to this token in learning will cause both weights to be increased. Even if the only token the learner has been exposed to is a [b], then, the learner will still be more likely to judge a novel voiced stop such as [g] as acceptable than a voiceless one such as [k], because attested [b] tokens are taken as support for both the specific sound [b] and for the wider class of voiced stops. This contrasts with specific-to-general models, which do not predict any generalization from an individual sound.

The predictions that the specific-to-general view makes are quite strong. If the specific-to-general assumption is combined with the assumption that the induction of sound-specific patterns presupposes a particular number of tokens of that sound (Adriaans & Kager, 2010), learning a pattern over a class of sounds requires at least as much exposure to the language as learning a pattern over a single sound; we test this prediction in Experiments 1 and 2a. Furthermore, a minimal generalization learner that has only been exposed to two words in a language, both of which start with the same sound, will conclude that all words in the language start with that sound: until shown evidence to the contrary, it will not generalize at all to other sounds. We test this prediction in Experiment 3.

1.1.1 The subset problem and indirect negative evidence

The specific-to-general assumption is a natural solution to the subset problem (Dell, 1981). This problem has been argued to affect learners that can only use positive evidence from attested forms, as is the case for human learners: a learner of English is rarely told explicitly that a particular form (e.g., **mpepm*) is phonotactically illegal. To illustrate the problem, suppose that the onsets that the learner has been exposed to are [b], [d] and [g]. This input is compatible with the following two grammars (among others): in Grammar 1, all words start with a voiced stop; in Grammar 2, words can start with any stop, either voiced or voiceless. The language generated by Grammar 1 is a subset of the language generated by Grammar 2. If at one point in the learning process the learner believes that Grammar 1 is correct, and later on encounters a word that starts with a voiceless stop (e.g., [k]), the learner can revise its decision and assume the less restrictive Grammar 2 instead. The reverse decision is *prima facie* impossible because of the absence of negative evidence: A learner that has chosen Grammar 2 would never receive evidence that the generalization was too wide.[2] To avoid overly broad generalizations, then, learners have to be conservative: "Whenever there are two competing grammars generating languages of which one is a proper subset of the other, the learning

strategy of the child is to select the less inclusive one" (Dell, 1981, p. 34). This strategy was later termed the Subset Principle (Berwick, 1985; Hale & Reiss, 2003).

The specific-to-general assumption clearly addresses the subset problem, but it is not the only solution to this problem. It is true that learners rarely receive direct evidence that certain sound combinations are impossible in their language;[3] however, they often do receive *indirect* negative evidence in the form of frequency asymmetries. Suppose that the learner is exposed to a language in which words start with either [b] or [d] (a simplified version of Experiment 1). After encountering two words in the language, one that starts with [b] and one that starts with [d], the learner might conclude that the best characterization of the phonotactics of the language is that words can start with any voiced stop. Yet as the learner encounters additional words, all of which starting with [b] and [d], the systematic absence of [g] onsets increasingly argues against the original generalization that any voiced stop can serve as an onset (Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007). This constitutes indirect negative evidence that may cause the learner to favor a narrower hypothesis, in this case that only [b] and [d] are legal onsets. Most probabilistic models will exhibit this behavior (Hayes & Wilson, 2008; Linzen & O'Donnell, 2015; Moreton et al., 2015).

In summary, then, unless additional assumptions are made, models that take into account indirect negative evidence predict that given frequency asymmetries suggesting that a gap in a generalization is systematic, the learner will conclude that the sound sequence missing from the input is essentially ungrammatical, and will not generalize to it. Models such as the minimal generalization learner, which is not sensitive to such frequency asymmetries, predict sustained generalization (for simulations, see Linzen & Gallagher, 2014; Linzen & O'Donnell, 2015). We contrast these predictions in Experiments 1 and 2a.

1.2 Overview of the experiments

This paper reports the results of four artificial language learning experiments. In Experiment 1, participants were taught a language in which all word onsets had the same voicing (all were voiced or all were voiceless). Participants were divided into four groups, each of which received a different amount of exposure to the language. After the exposure phase, participants judged novel test words for acceptability. These test words started with one of three types of onsets: onsets that the participants had encountered in the exposure phase (attested); new onsets that had the same voicing as the exposure onsets (conforming); and new onsets that had the opposite voicing from the exposure onsets (nonconforming). To anticipate the results, participants showed evidence of distinguishing voiced from voiceless onsets after very little exposure, but required more exposure to distinguish the onsets they were exposed to from the unattested but conforming onsets.

The language used in Experiment 1 had a categorical, phonetically based pattern. Experiment 2a tested the generality of the findings of Experiment 1 by teaching participants a language with an abstract phonotactic pattern that was not tied to a phonetic feature—specifically, identity between two consonants—using a similar paradigm. Further expanding on Experiment 1, the regularity in Experiment 2a was probabilistic rather than categorical. The results were qualitatively similar to those of Experiment 1: participants showed early distinction between test words with identical consonants and test words with non-identical consonants, followed by a gradually increasing preference for the exposure consonants.

In Experiment 2b, participants were taught a control language whose goal was to verify that the results

[2] In fact, under the assumption that simpler grammars—grammars that can be described more succinctly—are preferred to complicated ones (Chomsky & Halle, 1968), a learner would typically select the *widest* grammar possible, unless it is equipped with a countervailing bias such as the Subset Principle (Dell, 1981).

[3] This is a simplification: some phonotactic restrictions may be inferred from morphophonological alternations.

of Experiment 2a were indeed due to learning rather than a pre-existing preference for consonant repetition. Finally, Experiment 3 tested whether learners can generalize a phonotactic regularity to new sounds based on just a single instance of the regularity.

2 Experiment 1: A natural-class based generalization

The artificial language used in this experiment had a categorical natural-class based phonotactic regularity: all word onsets had the same voicing (either all voiced or all voiceless; different versions of the language were presented to different participants). Following the exposure phase, participants provided acceptability judgments for words of three types:

1. Conforming attested onset (CONF-ATT): words whose onset appeared as the onset of one or more of the exposure words. Since the phonotactic pattern was categorical, all of these onsets conformed to it.

2. Conforming novel onset (CONF-UNATT): words whose onset did not appear as the onset of any of the exposure words, but had the same voicing as those onsets.

3. Nonconforming unattested onset (NONCONF-UNATT): words whose onset differed in voicing from the onsets of all of the exposure words.

All of the exposure words differed from each other, and all of the test words were distinct from the exposure words. This was the case even for CONF-ATT test words: in that condition the onset was shared with some of the exposure words, but the full word was novel. Exposure sets were constructed which consisted of five words, one with each of the exposure onsets. Participants were divided into four groups; each group was given a different number of exposure sets (one, two, four or eight). For example, participants in the One Set group heard five exposure words, one with each of the exposure onsets, and participants in the Two Sets group heard ten exposure words, two with each of the exposure onsets. A detailed description is given in the Materials section below; see Table 1 for examples.

2.1 Method

2.1.1 Materials and procedure

The onsets of all of the stimuli used in the experiment were drawn from the set of six voiced obstruents [b d g ð v z] or from the set of six voiceless obstruents [p t k θ f s]. Words of the form $C_1V_1C_2V_2$ were created with all possible combinations of these onsets as the first consonant C_1; the vowels [a], [e], [i], [o] and [u] as V_1; the consonants [l], [m] and [n] as C_2; and the vowels [a], [i] and [u] as V_2. When the resulting combination formed an existing English word, one of the consonants [l], [m] or [n] was added to the end of the word (e.g., *tunal* instead of *tuna*).

The words of the language, as in all others languages used in this paper, were stressed on their first syllable. They were recorded by a native English speaker. 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.

Participants in each exposure group were assigned to one of 12 lists. All of the exposure words in each list had the same voicing: They were either all voiced or all voiceless. Five of the onsets were presented to the participants in exposure, and the sixth was held out. List 11, for instance, had exposure words with the onsets [p], [θ], [k], [f] and [t], but not [s]. Whether the exposure onsets were all voiced or all voiceless

Exposure	Test		
kelo	CONF-ATT	CONF-UNATT	NONCONF-UNATT
tanu	falu	soma	zila
fula	femi	sunu	zoma
θomi			
pinu			

(a)

Exposure		Test		
ganu	gomu	CONF-ATT	CONF-UNATT	NONCONF-UNATT
belu	bina	zini	dimu	talu
vimu	voni	zonu	dila	tumu
zela	zuli			
ðano	ðamu			

(b)

Table 1: Two examples of the materials presented to participants in Experiment 1. (a) One exposure set, voiceless exposure onsets, [s] held out; (b) Two exposure sets, voiced exposure onsets, [d] held out. Note that these are only two examples of the 48 possible lists (12 possible held out consonants, times four exposure groups).

ATT (attested): onset consonant (but not the full word) was encountered in exposure phase.
UNATT (unattested): onset consonant was not encountered in exposure phase.
CONF (conforming): onset consonant has the same voicing as the exposure words.
NONCONF (nonconforming): onset consonant has the opposite voicing from the exposure words.

was counterbalanced across participants, as was the identity of the held-out onset. For each list, one of the onsets showed in exposure was selected as the onset for the CONF-ATT test condition (e.g., for List 11 the CONF-ATT consonant was [f] as in *fumi*). The CONF-UNATT onset was the onset from the same voicing class as the exposure that was held out (in List 11, [s] as in *sona*), and the onset of the NONCONF-UNATT words was the consonant with the opposite voicing to the CONF-UNATT one (in List 11, [z] as in *zili*). Tables 1a and 1b illustrate the full set of materials in the one exposure and two exposures group respectively, each with a different counterbalancing list.

The list of exposure words was constructed in blocks, such that each consecutive block of five words had exactly one word starting with each of the five exposure onsets. Participants did not receive any indication of the structure of the lists. The order of onsets was pseudo-randomized within each block. Likewise, the segments selected for the V_1, C_2 and V_2 slots were pseudo-randomized in consecutive blocks such that each block contained all possible segments for the relevant slot. The test words were presented in two blocks of three tokens, one token for each of the onsets representing the CONF-ATT, CONF-UNATT and NONCONF-UNATT categories, in pseudo-random order (again without indication of the division into two blocks). The vowel pattern and medial consonants were randomized separately for each participant, such that the onsets were the only cue that systematically distinguished the test conditions.

2.2 Procedure

All experiments in this paper were conducted using Experigen, a JavaScript framework for running online experiments (Becker & Levine, 2010). Participants were recruited though Amazon Mechanical Turk. Results obtained using Mechanical Turk have been repeatedly shown to replicate established findings from the experimental behavioral research literature (Crump, McDonnell, & Gureckis, 2013). Participants were paid $0.65 for completing an experiment. They were told that they needed to be native speakers of English to complete the experiment. They were asked for their native language in a short demographic survey at the end of the experiment; data from participants who reported a native language other than English were removed. Participants were limited to those with IP addresses within the United States. We rejected participants who performed multiple experiments or multiple versions of the same experiment, and assigned the task to new participants to reach the intended sample size.

The experiments were split into an exposure phase and a test phase. In both phases, the words were presented in isolation—i.e., not in a continuous stream. Participants were told that the exposure phase would be followed by a test phase during which they will be required to decide if new words sounded like they could belong to the language they were listening to (for a similar task, see Moreton, 2008, 2012; Reeder, Newport, & Aslin, 2013). During the test phase, the instructions for the task were repeated after every test word. Only two answers were possible: “yes” and “no”.

2.2.1 Participants

Six participants completed each combination of the 12 lists and four exposure groups, for a total of 288 participants (72 participants per exposure group). Three participants were rejected because their reported native language was not English. We report data from the remaining 285 participants (116 women, 166 men, three unreported; median age: 30, age range: 18–68, one unreported).

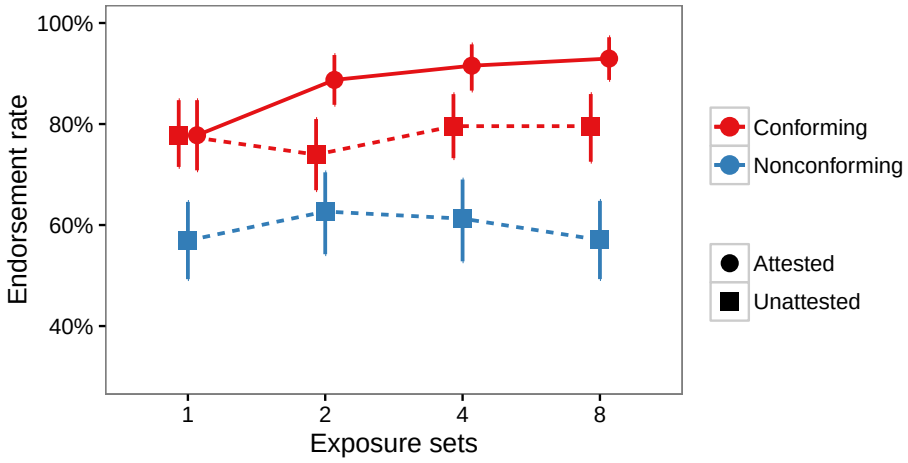

Figure 1: Mean endorsement rates for Experiment 1. Error bars represent bootstrapped 95% confidence intervals.

2.2.2 Statistical analysis

Logistic mixed-effects models (LMEM) (Baayen, Davidson, & Bates, 2008; Jaeger, 2008) were fitted to the participants' responses ("yes" or "no") using version 1.1.11 of the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015). There were only three test conditions, NONCONF-UNATT, CONF-UNATT and CONF-UNATT (there were no NONCONF-ATT test words since the phonotactic pattern held of all exposure words). As such, the design is not fully crossed, and we cannot estimate an interaction term between attestedness and conformity. We therefore treat the condition as a single three-level factor.

We fitted two types of models: full ANOVA models, which included all participants, and simple effect models, which only included participants in a given group (e.g., the Two Sets group). Fixed effects in the full models included group as a four-level factor and onset type as a three-level factor. All factors were coded using sum coding; the main effect of one factor can therefore be interpreted as its average effect across all levels of the other factors. The random effect structure for all models included a by-subject intercept and a by-subject slope for the effect of onset type, as well as a by-onset intercept. The statistical significance of each term in the model was assessed by comparing the likelihood of the full model to the likelihood of a model that did not include the factor in question, but did include the random by-subject slope for that factor as well as higher order interactions wherever applicable (Levy, 2014).

2.3 Results

The mean proportion of test words that participants in each group judged as acceptable in each of the conditions is shown in Figure 1. Visual examination of the results suggests that participants in all exposure groups distinguished CONF from NONCONF onsets in all exposure groups (except, perhaps, the Two Sets group); conversely, only the Two, Four and Eight Sets groups show a distinction between the two categories of CONF onsets (ATT and UNATT).

An ANOVA in the full factorial model, with all four exposure groups and three conditions, found a significant main effect of condition ($\chi^2(2) = 78.6$, $p < .001$); the main effect of exposure group did not reach significance ($\chi^2(3) = 7.7$, $p = .05$), but the interaction did ($\chi^2(6) = 13.6$, $p = .03$). Our main interest, however, is in the pairwise comparisons within the three levels of the Onset Type factor, which we

turn to next.

2.3.1 CONF-ATT vs. CONF-UNATT

An ANOVA including all exposure groups and only CONF-ATT and UNATT words found a significant effect of onset type ($\chi^2(1) = 18.4$, $p < .001$), such that CONF-ATT onsets were more likely to be endorsed than CONF-UNATT ones. The main effect of group was significant ($\chi^2(3) = 10.3$, $p = .02$), and so was the interaction between onset type and group ($\chi^2(3) = 9.9$, $p = .02$). Separate models fitted within each exposure group (simple effects) showed that this interaction was driven by the absence of a significant preference for CONF-ATT onsets in the One Set group ($\chi^2(1) = .94$, $p = .33$), compared with a significant preference for CONF-ATT onsets in the Two and Eight Sets groups and a nonsignificant preference in the Four Sets group (Two Sets: $\chi^2(1) = 8.1$, $p = .004$; Four Sets: $\chi^2(1) = 3.15$, $p = .08$; Eight Sets: $\chi^2(1) = 14.9$, $p < .001$).

How much support do the results of the One Set group results provide for the "null" hypothesis, according to which there is no difference between attested and unattested CONF onsets? We calculated the appropriate Bayes factor using the Bayes Information Criteria approximation (Kass & Raftery, 1995; Wagenmakers, 2007); the result was 10.6, corresponding to a posterior probability of approximately 91% for the null hypothesis assuming a uniform prior. This Bayes factor is characterized by Wagenmakers (2007) as indicating "positive" evidence for the null hypothesis.

2.3.2 NONCONF-UNATT vs. CONF-UNATT

CONF-UNATT words were more likely to be endorsed than NONCONF-UNATT ones ($\chi^2(1) = 32.0$, $p < .001$). The effect of group was not significant ($\chi^2(3) = 0.6$, $p = .9$) and neither was the interaction between group and type ($\chi^2(3) = 2.85$, $p = .41$). Within-group models showed that the effect reached significance for all groups except for the Two Sets group, where the effect was in the same direction as in the rest of the groups but was not significant (One Set: $\chi^2(1) = 12.0$, $p < .001$; Two Sets: $\chi^2(1) = 2.56$, $p = .11$; Four Sets: $\chi^2(1) = 8.08$, $p = .004$; Eight Sets: $\chi^2(1) = 11.6$, $p < .001$).

2.3.3 Differences across counterbalancing lists

As mentioned in the Materials section, the voicing of the onsets of the exposure words was counterbalanced across participants, as was the identity of the held-out consonant. This resulted in 12 lists in total. As a post-hoc analysis, we explored whether there were differences across those lists. Since there were only six subjects in each combination of list and exposure group, we pooled together the lists based on the voicing and manner of articulation of the held-out consonant – e.g., the lists where [p], [t] and [k] were the held-out consonants are collapsed into a single voiceless stop category. Figure 2 plots the results broken down in this way. Differences across the lists appear to be minor, although the high uncertainty (due to the low number of trials) makes it difficult to draw definite conclusions (for example, there appears to be a tendency for the One Set group to distinguish attested from unattested onsets in voiced stop lists).

We repeated the statistical comparison between CONF-ATT and CONF-UNATT words in two ways. First, we added a manner factor, indicating whether the held-out consonant was a stop or a fricative, as well as the interaction of that factor with condition. The effects of these predictors were not significant in any of the exposure groups. Second, we added a voicing factor and its interaction with condition. The main effect of voicing and the interaction were not significant in the One, Two and Eight Sets groups. There was a significant interaction in the Four Sets group, such that the effect of condition was larger when the exposure

onsets were voiceless ($\chi^2(1) = 7.57$, $p = .006$). Since a result restricted to the Four Sets group does not have a clear interpretation, and this was one of a large number of post-hoc tests, we do not comment on this finding any further. A higher-powered investigation of this difference may be an interesting direction for future research.

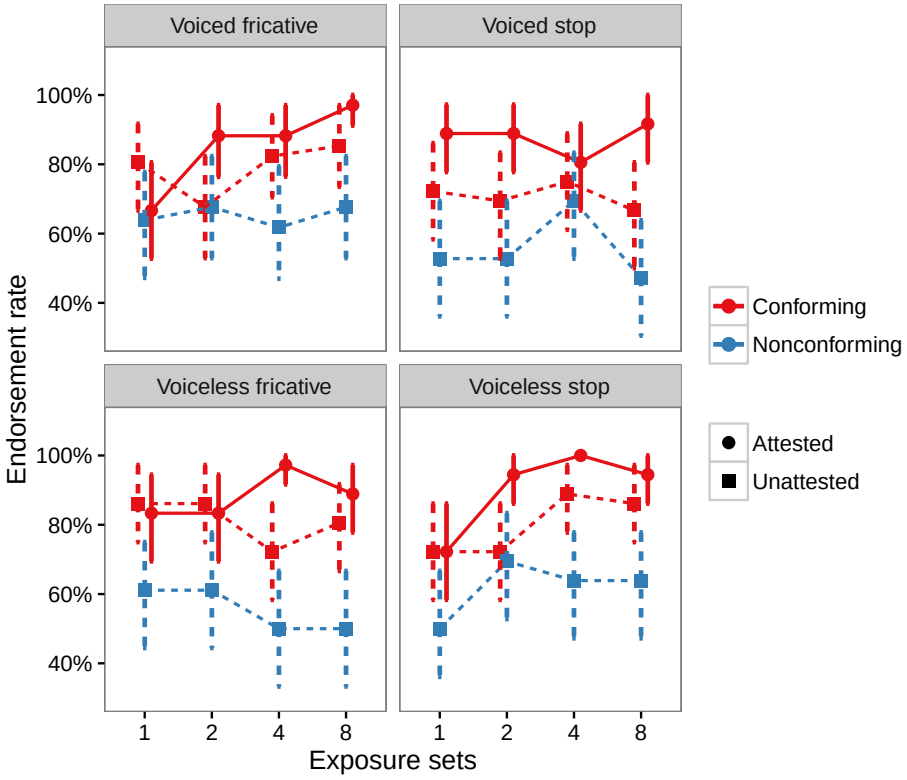

Figure 2: Endorsement rates for Experiment 1, broken down by the voicing and manner of articulation of the held out (CONF-UNATT) onset. Error bars represent bootstrapped 95% confidence intervals.

2.4 Discussion

Participants in Experiment 1 were taught artificial languages that had a categorical natural-class based phonotactic regularity: all word onsets had the same voicing (either all voiced or all voiceless, depending on the list). Participants then judged the acceptability of novel words with onsets of three types: CONF-ATT onsets, which were encountered during exposure; CONF-UNATT onsets, which shared the value for the voicing feature with the onsets of the exposure words but were not encountered during exposure; and NONCONF-UNATT onsets, which had the opposite value for the voicing feature than the exposure words.

CONF-UNATT onsets were consistently endorsed more often than NONCONF-UNATT onsets, regardless of the amount of exposure: even after a single set of exposure to each onset type, participants preferred onsets with the same voicing as the exposure onsets to onsets with the opposite voicing. Conversely, participants did not start distinguishing CONF-ATT from CONF-UNATT onsets until after two or more exposure sets. The three-way distinction between CONF-ATT, CONF-UNATT and NONCONF-UNATT words was similar in the Two Set, Four Set and Eight Set groups: despite growing indirect negative evidence suggesting that not all

conforming onsets occur in the language, participants continued to generalize beyond the attested onsets.

Words that started with a NONCONF-UNATT onset were judged to be acceptable at a fairly high rate (around 60% of the time), even after eight exposure sets. This is likely to reflect the fact that onset voicing is far from the only possible dimension of generalization from the exposure words. Just as all exposure words had the same voicing, they also all started with a consonant, had two syllables, were stressed on their first syllable, and so on. We suspect that test words that differed from the exposure words in more dimensions, such as *ulpiuzi* or *eh*, would have been endorsed at a lower rate. The results of the current experiment do not allow us to delineate exactly how far participants would be willing to generalize: the only conclusion we can be confident about is that either they did not generalize to NONCONF-UNATT onsets at all, or if they did, they did so to a lesser extent than to CONF-UNATT onsets.

3 Experiment 2a: A probabilistic abstract generalization

Participants in Experiment 1 generalized rapidly, before they were able to distinguish the sounds they were exposed to from unattested but similar sounds. They continued to generalize even after as many as eight exposure sets. Experiment 2a tests the generality of that result by applying the same experimental paradigm to a language that differs from the language of Experiment 1 in two ways.

First, generalization in Experiment 1 was supported by a categorical regularity: all of the words in the language had the same voicing. There is evidence that speakers' knowledge of the distribution of sounds in their language is not limited to the categorical distinction between possible and impossible sound sequences; rather, speakers keep track of the relative frequencies of the possible sounds and sound sequences (Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997). For example, although neither of the nonwords *riss* [ɹɪs] or *youdge* [jaʊdʒ] contains any sounds or sub-sequences of sounds that are unattested in English,[4] the nonword *riss*, which is comprised of frequent sound sequences, is judged to be a more likely potential word of English than *youdge* (Coleman & Pierrehumbert, 1997; Jusczyk & Luce, 1994).

Second, the generalization in Experiment 1 was stated over a phonetically defined class of sounds. While many phonotactic generalizations in natural language are based on the phonetic properties of individual sounds, some generalizations are abstract, in that they do not make reference to the phonetic properties of any particular sound. The simplest type of such a generalization is sound identity (repetition). A large range of studies have shown that such generalizations can be acquired in artificial language studies (Gervain, 2014; Marcus et al., 1999; Moreton, 2012). In natural languages, generalizations that have to do with segment repetition or bans on repetition have been documented in Yucatec Mayan, Hebrew, Peruvian Aymara and other languages (Berent, Marcus, Shimron, & Gafos, 2002; Gallagher, 2013).

To replicate the findings of Experiment 1 and broaden the scope of the conclusions that can be drawn from those findings, then, Experiment 2a tested whether the pattern of results held for a probabilistic abstract generalization. All of the words in the language used in this experiment had the form $C_1V_1C_2V_2$ (e.g., *semi*). Vowels in the language varied freely, and the consonant pairs followed one of eight consonant-specific phonotactic regularities. Four of those regularities involved two different consonants, e.g., C_1 = [k] and C_2 = [s]; two words conforming to this particular regularity are *kesa* and *kisu*. The other four involved two particular identical consonants, e.g., C_1 = [p] and C_2 = [p] (as in *pepu*).

While the phonotactics of the language can be captured precisely using these eight consonant-pair spe-

[4]Consider the English words *yowl* [jaʊl] ('a loud wailing cry') and *gouge* [gaʊdʒ] ('a chisel with a concave blade').

Exposure	Test	
CONF	CONF-ATT	CONF-UNATT
pipa	pepu	kuka
ʃuʃe	ʃaʃi	sesi
gagu	gugi	dʒidʒe
nuni	ninu	mamu
NONCONF	NONCONF-ATT	NONCONF-UNATT
kesa	kasi	pina
mudʒe	medʒa	nage
dʒeku	dʒuke	gaʃe
sami	sime	ʃipu

Table 2: Illustration of the materials presented to the participants in Experiment 2a. The table shows a complete exposure and test set for the One Set group (with one possible set of vowel patterns).

cific regularities, it was also the case that half of the words in the language followed the abstract regularity $C_1 = C_2$, much more than would be expected by chance. If participants learned this abstract generalization, they should generalize it to words that contain identical consonants outside of those included in the exposure phase. As mentioned above, numerous studies have shown that participants are able to learn repetition patterns (Endress & Bonatti, 2007; Gerken, 2006; Gervain, 2014; Marcus et al., 1999); our goal is to build on those studies to investigate how generalization to new repeated consonants depends on the amount of exposure to the language.

As in Experiment 1, exposure sets were created that included exactly one word that followed each of the narrow regularities, for a total of eight words per exposure set (see Table 2). The language was taught to several groups of participants, each receiving a different number of exposure sets. In the test phase, participants were asked to judge the acceptability of new words that had either consonant pairs that were familiar from the exposure phase (ATT) or new consonant pairs (UNATT). Half of the new consonant pairs had identical consonants (CONF) and half had non-identical consonants (NONCONF). In contrast with Experiment 1, the fact that only half of the exposure words followed the repetition regularity made it possible to construct NONCONF-ATT words. This led to a fully crossed design that allowed us to test for the independent contribution of the broad regularity (two identical consonants) and narrow regularities (the first consonant is [k] and the second consonant is [s]; the first consonant is [p] and the second is also [p]; etc.).

3.1 Method

3.1.1 Materials and procedure

All words in the experiment were of the form $C_1V_1C_2V_2$, e.g., *kesa*. The exposure words had one of eight different consonant pairs, four of which were identical and four of which were not (see Table 2). All participants were presented with 16 testing words, eight with the consonant pairs heard in exposure (ATTESTED) and eight with new consonant pairs (UNATTESTED). Each of the individual consonants C_1 and C_2 in the unattested consonant pairs were encountered during the exposure phase, in both initial and medial position, but not as a combination. 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 and V_2; e.g., for [p p], the words constructed were *pipa*, *pipe*, *pupa* and so on. The stimuli were recorded by a female native English speaker.

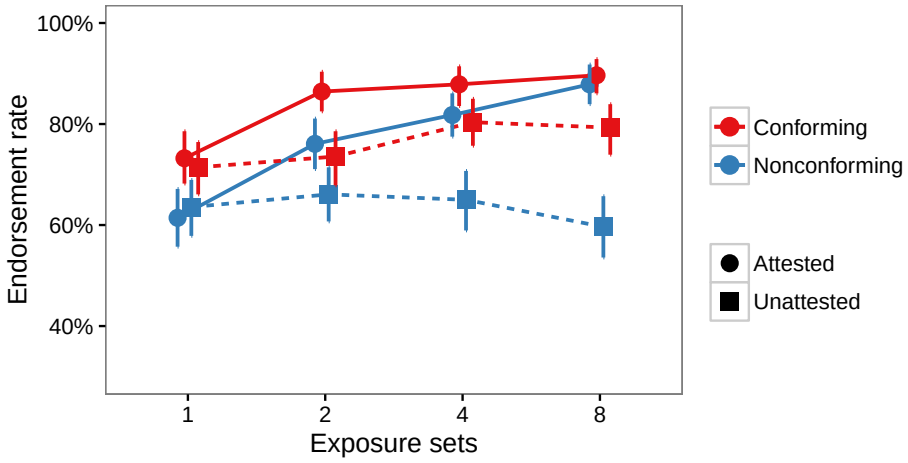

Figure 3: Mean endorsement rates for Experiment 2a. Error bars represent bootstrapped 95% confidence intervals.

In the exposure phase, participants listened to one, two, four or eight exposure sets. All exposure words differed from each other; that is, the same consonant pair was never heard with the same vowels more than once. As in Experiment 1, the specific words from the exposure phase were never repeated in the test phase. For example, if *bagu* and *biga* appeared in the exposure phase, neither could appear in the test phase, but *bega* could.

All participants were exposed to the same C_1-C_2 pairs, though the particular words (i.e., the combinations of consonant pair and vowel patterns) differed across participants. Items were pseudo-randomized in blocks as in Experiment 1. In particular, the vowel patterns were randomized for each participant separately, such that the consonant pair was the only cue that systematically distinguished the test conditions from each other.

3.1.2 Participants

A total of 280 participants completed the experiment, 70 in each group. Demographic information was not collected due to a technical failure.

3.1.3 Statistical analysis

As in Experiment 1, we fitted a full model that included participants from all four groups, as well as within-group models for each of the groups. The full model had three fixed effects: one between subjects (the exposure group) and two within subjects (Attestation and Conformity). The random effect structure for subjects in the full model included an intercept and random slopes for Attestation, Conformity and the interaction between the Attestation and Conformity; we also had a random intercept for the consonant pair. As before, p-values were calculated using the chi-square approximation to the likelihood ratio test.

3.2 Results

3.2.1 Full model

Figure 3 illustrates the mean endorsement rates for each group and condition. The full statistical model yielded an effect of group ($\chi^2(3) = 25.6$, $p < .001$), reflecting the fact that endorsement rates were higher on average for participants who received more exposure to the language. There was also an effect of Attestation, reflecting higher average endorsement rates for words with ATT than for words with UNATT consonants ($\chi^2(1) = 11.7$, $p < .001$), and an effect of Conformity, reflecting higher average endorsement rates for CONF than for NONCONF words ($\chi^2(1) = 11.1$, $p < .001$).

The effect of Attestation was modulated by an interaction with group ($\chi^2(3) = 35.6$, $p < .001$), which reflects the fact that participants were better at distinguishing ATT from UNATT items the more exposure they received to the language. The interaction of group and Conformity was not significant ($\chi^2(3) = 1.3$, $p = .73$), and neither was the interaction between Conformity and Attestation ($\chi^2(1) = .04$, $p = .85$). The interpretation of these findings is complicated by the significant three-way interaction ($\chi^2(3) = 8.44$, $p = .04$); Figure 3 suggests that the three-way interaction reflects the fact that as participants received additional exposure sets, the effect of Conformity gradually diminished, but only for test words with ATT consonants; the effect of Conformity was robust for test words with UNATT consonants even in the Eight Sets group.

3.2.2 Within-group models

One Set: In this group, CONF test words were endorsed significantly more often than NONCONF ones ($\chi^2(1) = 9.4$, $p = .002$). The effect of Attestation and the interaction did not reach significance (Attestation: $\chi^2(1) = .02$, $p = .9$; interaction: $\chi^2(1) = .5$, $p = .47$), suggesting that the narrow phonotactic regularities did not affect endorsement rates (the numerical endorsement rates were: CONF-ATT: 73%; CONF-UNATT: 71%; NONCONF-ATT: 61%; NONCONF-UNATT: 64%).

The Bayes factor in support of the null hypothesis of no Attestation main effect was 33.2. A similar test for the interaction yielded a Bayes factor of 25.7. Both values are characterized as providing "strong" evidence for the null hypothesis (Wagenmakers, 2007).

Two Sets: Test words with ATT consonants were judged to be acceptable significantly more often than ones with UNATT consonants ($\chi^2(1) = 11.6$, $p < .001$). The effect of Conformity was also significant ($\chi^2(1) = 7$, $p = .008$), and the interaction was not ($\chi^2(1) = 1.6$, $p = .21$).

Four Sets: Both of the effects of the factors reached significance; the interaction was again nonsignificant (Conformity: $\chi^2(1) = 9.4$, $p = .002$; Attestation: $\chi^2(1) = 8.5$, $p = .004$; interaction: $\chi^2(1) = .6$, $p = .26$).

Eight Sets: Again, both main effects were significant (Attestation: $\chi^2(1) = 17$, $p < .001$; Conformity: $\chi^2(1) = 9.2$, $p = .002$). In contrast with the Two Sets and Four Sets condition, the interaction was also significant ($\chi^2(1) = 4.28$, $p = .04$). Consistent with the interaction, separate models fitted within ATT and UNATT items (both with by-subject Conformity slopes) found that Conformity had a significant effect for UNATT items ($\chi^2(1) = 17.6$, $p < .001$) but no discernible effect for ATT ones ($\chi^2(1) = .34$, $p = .56$).

3.3 Discussion

After a single exposure to each of the eight possible consonant pairs, four 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 the particular pair of identical consonants was presented in the exposure phase. Participants did not start showing evidence of having learned individual consonant pairs until they received at least two exposure sets (i.e., two words with each consonant pair).

As in Experiment 1, participants consistently generalized to CONF-UNATT words even after eight exposure sets. To further explore this sustained generalization pattern, we administered the experiment to an additional group of 70 participants, this time with 16 exposure sets. Since we only had 12 distinct words with each consonant pair, some of the exposure words were repeated twice. It was still the case, however, that none of the test words occurred in the exposure phase.

The endorsement rates for the 16 Sets group were similar to the ones for the Eight Sets group, with the exception that the endorsement rate for NONCONF-UNATT words was more similar to the endorsement rate for those words in the One, Two and Four groups (CONF-ATT: 92%; CONF-UNATT: 79%; NONCONF-ATT: 89%; NONCONF-UNATT: 67%); this suggests that the dip in endorsement rates for NONCONF-UNATT in the Eight Sets group visible in Figure 3 was spurious. The two main effects were significant (Attestation: $\chi^2(1) = 29.2$, $p < .001$; Conformity: $\chi^2(1) = 8.8$, $p = .003$), but the interaction was not ($\chi^2(1) = .7$, $p = .41$; all models were fitted without a correlation term between the by-subject intercept and slopes due to model convergence issues). The simple effect of Conformity was significant within UNATT words ($\chi^2(1) = 4.77$, $p = .03$) but not within ATT ones ($\chi^2(1) = .05$, $p = .83$). In sum, statistical evidence for generalization to CONF-UNATT words remains robust even for participants who received 16 exposure sets; the fact that this evidence was weaker than in the Eight Sets group may be an artifact of spuriously low endorsement rates for NONCONF-UNATT words in the Eight Sets group.

In conclusion, participants generalized to unattested consonant pairs after very little exposure to the language, and continued to generalize even after being given ample indirect negative evidence suggesting that only certain consonant pairs can appear in the language.

4 Experiment 2b: Ruling out a pre-existing preference for identity

We interpreted our participants' preference for words with repeated consonants in the One Set group of Experiment 2a as reflecting rapid phonotactic generalization. Before being confident in this interpretation, however, we must rule out the possibility that the higher endorsement rate for test items with identical consonants was due to a prior preference for words with identical consonants rather than due to exposure to the artificial language. Such a prior preference could be derived from the participants' native language or from any number of perceptual or cognitive factors (Endress & Bonatti, 2007; Gervain, 2014).

Experiment 2b was designed to test for such a pre-existing preference for words with identical consonants. Participants were exposed to eight words, each containing a different non-identical consonant pair. After the exposure phase, participants judged the unattested items from the test phase of Experiment 2a (both CONF-UNATT and NONCONF-UNATT). An outcome in which participants still showed a preference for identical over non-identical items despite not having seen any identical items in exposure would be consistent with a pre-existing preference for identical items. If, on the other hand, participants showed no identity preference in the test phase, the interpretation of the identity preference in Experiment 2a as being due to learning would stand.

Exposure	Test	
	NONCONF-ATT	CONF-UNATT
ʃidʒa	ʃadʒi	keku
mune	mene	sasi
sagu	sugi	dʒidʒe
pusi	pisu	mamu
geka	gaki	
kupe	kepa	NONCONF-UNATT
nuʃe	naʃe	ʃipu
dʒami	dʒima	pina
		nage
		gaʃe

Table 3: All consonant pairs used in exposure and test for Experiment 2b, with randomly selected example words.

4.1 Method

4.1.1 Materials and procedure

All words had the form $C_1V_1C_2V_2$, as in Experiment 2a. As in the One Set group of Experiment 2a, there were eight exposure words and 16 test words. All exposure words had two non-identical consonants (see Table 3). Vowel patterns were chosen at random, with no vowel pattern repeated across the exposure and test words. As in Experiment 2a, half of the test words had consonant pairs encountered in exposure (ATTESTED) and half did not (UNATTESTED). All of the test words in the ATT condition had non-identical consonants encountered in the exposure phase. The unattested words in testing had the same consonant pairs as in Experiment 2a, half identical and half non-identical (four of each). For consistency with Experiment 2a, we still use the labels CONF and NONCONF to refer to the test words with identical and non-identical consonants respectively, even though the exposure phase in Experiment 2b did not provide any evidence for the segment-identity generalization. Since no exposure words had identical consonants, there were no CONF-ATT test items; the three test conditions were NONCONF-ATT, CONF-UNATT and NONCONF-UNATT.

The support that CONF and NONCONF test words received from irrelevant natural-class based patterns in the exposure set was matched as follows. Each of the eight consonants in the language appeared in the exposure phase once in initial position and once in medial position. As such, the CONF-UNATT and NONCONF-UNATT test words received equal support from the positional frequency of the individual consonants, as in Experiment 2a. In addition, CONF-UNATT and NONCONF-UNATT test words were matched for the amount of natural class based support they received from consonant co-occurrences in the exposure word (voicing, place of articulation and manner of articulation). For example, the test word with the consonants [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 exposure set, so its total natural class-based co-occurrence support score is 2. It is matched with [g]–[ʃ], which also receives natural-class based support from two 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 exposure set.

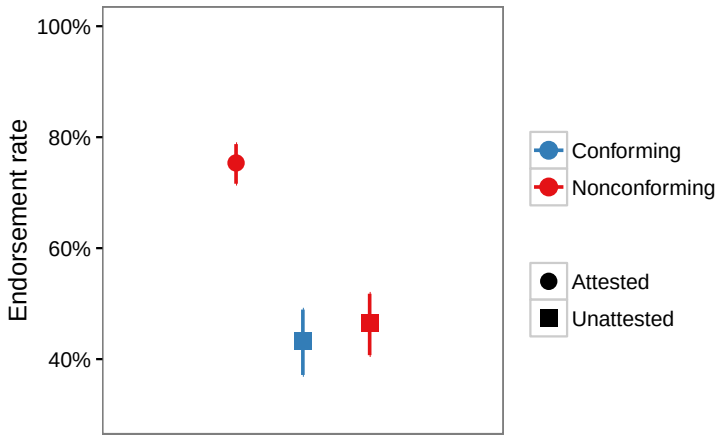

Figure 4: Mean endorsement rates for Experiment 2b. Error bars represent bootstrapped 95% confidence intervals.

4.1.2 Participants

A total of 70 participants completed the experiment (34 women, 35 men, one unreported; median age: 27, age range: 18-61).

4.1.3 Statistical analysis

A LMEM was fitted to the results, with a three-level factor of consonant type (NONCONF-ATT, CONF-ATT, CONF-UNATT) as a fixed effect, as well as random intercepts for consonant pairs and subjects and a random slope by subject for consonant type.

4.2 Results

The results of Experiment 2b are shown in Figure 4. Contrary to the predictions of the pre-existing preference hypothesis, participants did not show a preference for CONF-UNATT words; if anything, there was a slight preference for NONCONF-UNATT words over CONF-UNATT ones. There was a striking difference between NONCONF-ATT words and both CONF-UNATT and NONCONF-UNATT words: unlike the One Set group of Experiment 2a, participants in Experiment 2b were much more likely to endorse test words with attested than unattested consonant pairs.

Statistical analysis showed that the effect of condition on endorsement rates was significant ($\chi^2(2) = 27.6$, $p < .001$). We performed planned comparisons to examine the difference between the different levels of the factor. In line with Figure 4, the difference between NONCONF-ATT on the one hand and CONF-UNATT and NONCONF-UNATT on the other hand (i.e., the two UNATT conditions collapsed together) was significant ($\chi^2(1) = 27.4$, $p < .001$). By contrast, the difference between CONF-UNATT and NONCONF-UNATT did not approach significance ($\chi^2(1) = .57$, $p = .45$).

4.3 Discussion

Participants in Experiment 2b, who were not exposed to identical consonant pairs, did not show any preference for novel items with identical consonants (CONF-UNATT). The results therefore support the learning

hypothesis, according to which the preference for identical items after one exposure in Experiment 2a was due to learning during the experiment. Thus, our interpretation of the results of the One Set group in Experiment 2a remains unchanged.

The results of Experiment 2b reveal an additional effect. Unlike in Experiment 2a, participants in Experiment 2b showed a strong preference for attested over unattested consonant pairs after just one exposure. While we cannot make firm claims about the source of this difference, one possibility is that the presence of a broad regularity interferes with the learning of narrower regularities. In Experiment 2a, the presence of the identity regularity may have prevented learners from attending sufficiently to the narrower regularities with small amounts of exposure, while in Experiment 2b learners were free to focus on the specific, attested consonant pairs.

At first blush, the lack of a preference for identical items in Experiment 2b compared to Experiment 2a could still be consistent with a pre-existing preference for identical items: The absence of identical consonant pairs from the exposure data could have been taken as evidence for the generalization that pairs of identical consonants are underattested, offsetting a pre-existing preference for identical consonants. However, this alternative explanation for the results of Experiment 2b becomes less plausible if we consider the radically different amount of support for the generalization that the exposure 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 2a, on the other hand, the participants received four identical pairs instead of the expected one pair, for an O/E of $4/1$. In other words, the evidence for the overattestation of identical pairs in Experiment 2a is much stronger than the evidence for their underattestation in Experiment 2b. It is therefore implausible to assume that the preference for identical items after one exposure in Experiment 2a was due to pre-existing preference, and at the same time that the lack of preference for identical items in Experiment 2b was due to learning during the experiment that offset that preference.

5 Experiment 3: Generalization from a single type

Participants in Experiments 1 and 2a showed evidence of rapid phonotactic generalization. That evidence preceded any evidence that they had learned the narrower, sound-specific phonotactic patterns (i.e., that [k] is an allowed onset). What is the minimal amount of evidence that is required for participants to begin generalizing? In particular, would they generalize based on a *single* type of onset consonant, or would they wait until they have encountered multiple examples of a phonological class before they begin generalizing to other members of that class, as argued by the minimal generalization hypothesis (Albright & Hayes, 2003; Albright, 2009; Adriaans & Kager, 2010)? Experiment 3 addresses this question by exposing participants to a language in which a particular dimension of generalization is only supported by a single type of sound. If participants still generalized along that dimension, the conclusion would be that learners can generalize based on a single type.

The exposure set in the critical group in Experiment 3 contained only one type of voiceless stop onset ([p], [t] or [k], counterbalanced across participants). Participants were tested to determine if they endorsed the two voiceless stops they had not encountered in the exposure phase; for example, if [p] was the voiceless stop encountered in the exposure phase, the generalization onsets were [k] and [t]. Only two words starting with the voiceless stop were presented in the exposure phase. Six words starting with approximant onsets ([l], [w] or [y]) were added to the exposure set to make the training phase longer. As in Experiment 1,

participants judged three kinds of test items: CONF-ATT, CONF-UNATT and NONCONF-UNATT. We refer to this language as the Single Type language.

The experiment included two additional languages designed to allow us to draw firmer conclusions from the findings related to the Single Type language. The Two Types language included two different voiceless stops in the exposure set, e.g., [t] and [k]. One word starting with each onset was presented in the exposure phase. Based on the results of Experiment 1, we expect participants assigned to the Two Types language to generalize to the unattested voiceless stop, and to fail to distinguish attested from unattested voiceless stops. Finally, the Control language did not include any voiceless stops at all: participants who were assigned this language were only exposed to the six approximants. This language served to determine whether participants had a pre-existing bias for or against voiceless stop onsets.

5.1 Method

5.1.1 Materials

Words were created with three classes of onsets: voiceless stops ([p], [t] and [k]), which we refer to as CONF onsets; voiced fricatives ([z] and [ð]), which we refer to as NONCONF onsets; and approximants ([w], [y] and [l]), which we refer to as APPROX onsets. All onsets were embedded in words of the form $C_1V_1C_2V_2$, where the medial consonant C_2 was one of the nasals [m] or [n], and the vowel pattern V_1–V_2 was one of [a]–[i], [a]–[i], [u]–[a] or [i]–[a]. All possible combinations of onset, medial consonant and vowel pattern were recorded by a male native English speaker.

Participants were divided into three groups. Each group was assigned to one of the languages (Control, Single Type or Two Types). The exposure phase in all languages included six APPROX words, two starting with each of the onsets [w], [y] and [l]. Participants who were taught the Control language were only exposed to the APPROX words (see Table 4c). The Single Type language additionally included two words starting with the same CONF onset ([p], [t] or [k], counterbalanced across participants; see Table 4a). Finally, the exposure phase in the Two Types language included two words, each starting with a different CONF onset ([p] and [t], [p] and [k], or [t] and [k], counterbalanced across participants), in addition to APPROX words (see Table 4b). All participants received a single exposure set.

The approximants [w], [y] and [l] are considered to be voiced consonants that are neither stops nor fricatives (Hayes, 2011). If anything, these onsets should provide support for the voiced fricative test onsets (NONCONF-UNATT) rather than the voiceless stop ones (CONF-ATT). Any preference for CONF-UNATT over NONCONF-UNATT test onsets, then, would be observed despite rather than because of the APPROX onsets.

In the test phase, all participants judged five novel words, one with each of the five onsets [p], [t], [k], [z] and [ð]. For consistency, we refer to [p], [t] and [k] as CONF test onsets and to [z] and [ð] as NONCONF test onsets in all three languages, even though one of them, the Control language, did not provide any basis for generalizing to voiceless stops. None of the languages had NONCONF onsets in the exposure phase; in other words, NONCONF onsets were always unattested (NONCONF-UNATT). The exposure phase of the Control language did not have any CONF onsets; [p], [t] and [k] were therefore all CONF-UNATT. The test phase of the Single Type language had one CONF-ATT and two CONF-UNATT onsets, and the test phase of the Two Types language had two CONF-ATT and one CONF-UNATT onsets.

5.1.2 Participants

A total of 450 participants were recruited through Amazon Mechanical Turk: 50 participants in each of the three lists for the Single Type and Two Types languages, and 150 participants in the Control language. Nine

Exposure		Test		
APPROX	CONF	CONF-ATT	CONF-UNATT	NONCONF-UNATT
wamu	kami	kuna	pami	ðima
yuna	kamu		tanu	zanu
lani				
wina				
yani				
lima				

(a)

Exposure		Test		
APPROX	CONF	CONF-ATT	CONF-UNATT	NONCONF-UNATT
wamu	kami	kuna	pami	ðima
yuna	tamu	tanu		zanu
lani				
wina				
yani				
lima				

(b)

Exposure	Test	
APPROX	NONCONF-UNATT	CONF-UNATT
wamu	ðima	kuna
yuna	zanu	pami
lani		tanu
wina		
yani		
lima		

(c)

Table 4: Example of materials used in Experiment 3. (a) Single Type language, in the list that had [k] as the exposure CONF onset; (b) Two Type language, in the list the had [k] and [t] as the exposure CONF onsets; (c) Control language.

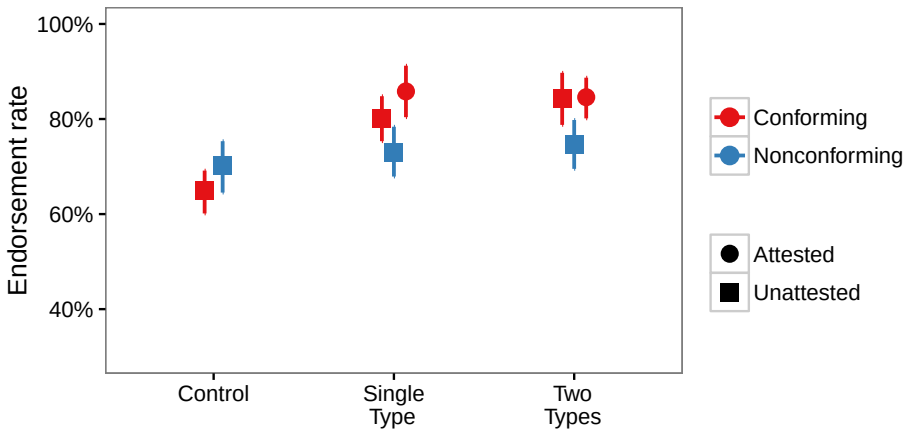

Figure 5: Mean endorsement rates for Experiment 3. Error bars represent bootstrapped 95% confidence intervals.

participants were rejected because they reported that English was not their only native language. We report data from the remaining 441 participants (233 women, 204 men, four unreported; median age: 28, age range: 18-71, one unreported).

5.1.3 Statistical analysis

The statistical analysis was similar to the previous experiments, with the exception that our design did not allow us include an onset type random slope for participants, since we only had a single observation per participant for some of the combinations of onset category and language (e.g., there was only one test token with a CONF-UNATT onset in the Two Types language). As such, the random effect structure in all LMEMs reported below only included random intercepts for subjects and for onsets.

5.2 Results

Figure 5 shows the mean endorsement rates for each onset type in each of the languages. The design was not fully crossed due to the absence of CONF-ATT onsets from the test phase of the Control language. Consequently, we performed two separate analyses: one that included all three languages, but only test words with CONF-UNATT and NONCONF-UNATT onsets; and another that included all three onset types, but only the Single Type and Two Types languages.

5.2.1 CONF-UNATT vs. NONCONF-UNATT onsets

Within these two conditions, the main effect of onset type was significant ($\chi^2(1) = 5.4$, $p = .02$), but not the main effect of language ($\chi^2(2) = 1.8$, $p = .4$). The interaction between onset type and language was significant ($\chi^2(2) = 12$, $p = .002$). This interaction was driven by higher endorsement rates for CONF-UNATT than NONCONF-UNATT onsets in both the Single Type and Two Types languages (Single Type: $\chi^2(1) = 4.11$, $p = .04$; Two Types: $\chi^2(1) = 5.99$, $p = .01$), but not in the Control language, where there was a nonsignificant difference in the opposite direction ($\chi^2(1) = 2.68$, $p = .1$).

The significant simple effect in the Single Type language suggests that learners generalized based on a single CONF onset type in exposure. The nonsignificant difference in the opposite direction in the Control language may reflect a tendency to interpret the approximant APPROX onsets in exposure as providing support for voiced over voiceless onsets.

5.2.2 Excluding the Control language

The effect of onset category was significant ($\chi^2(2) = 12.8$, $p = .002$); the effect of language was not significant ($\chi^2(1) = 0.13$, $p = .72$), and neither was the interaction ($\chi^2(2) = 1.22$, $p = .54$). This indicates that the pattern of results is not statistically different across the Single Type and Two Types language.

To further examine the effect of onset category, we performed pairwise comparisons across the levels of this factor. The difference in endorsement rate between CONF-UNATT and NONCONF-UNATT was significant ($\chi^2(1) = 8.3$, $p = .004$) and did not interact with language ($\chi^2(1) = .4$, $p = .53$). There was no significant difference between test words with CONF-ATT and CONF-UNATT onsets ($\chi^2(1) = 1.5$, $p = .21$), and again no interaction with language ($\chi^2(1) = 1.2$, $p = .27$).

Finally, we assessed the statistical significance of the difference between words with CONF-ATT and CONF-UNATT onsets within each language separately. Endorsement rates within the Two Types language did not differ across these conditions ($\chi^2(1) = 0$, $p = .96$); the numerical preference for CONF-ATT over CONF-UNATT onsets in the Single Type language did not reach significance ($\chi^2(1) = 2.79$, $p = .09$).

5.2.3 Differences across counterbalancing lists

As mentioned above, the voiceless consonant presented in the exposure phase in the Single Type language was [p], [t] or [k], counterbalanced across participants. As a post-hoc analysis, we explore whether the identity of the voiceless consonant in exposure affected participants' generalization patterns. We plot the endorsement rates in the Single Type language broken down by exposure consonant in Figure 6. The most salient pattern is that the difference between CONF-UNATT and CONF-ATT is clearer when [k] is the exposure consonant than when it is [p] or [t].

We next fit a mixed-effects logit model to the results of the Single Type language. The fixed effects were condition (CONF-UNATT, CONF-ATT and NONCONF-UNATT), exposure consonant ([p], [t] or [k]) and their interaction. We additionally had random subject and onset intercepts. There was a significant main effect of condition ($\chi^2(2) = 6.2$, $p = .04$), but the main effect of exposure consonant and the interaction were not significant (exposure consonant: $\chi^2(2) = 1.6$, $p = .44$; interaction: $\chi^2(4) = 7.2$, $p = .13$). From a statistical point of view, there is no clear evidence of a difference across the counterbalancing lists; at the same time, it is clear that the effect of condition in the Single Type language is primarily due to the group of participants that were exposed to [k]-initial words.

5.3 Discussion

What are the limits of rapid phonotactic generalization? The minimal generalization hypothesis (Adriaans & Kager, 2010; Albright, 2009) argues that learners need to be exposed to multiple types exemplifying a phonotactic pattern before they can generalize the pattern to new sounds. Experiment 3 tested this hypothesis by exposing participants to the Single Type language, in which two tokens of a single type of voiceless stop onset—e.g., [p]—were the only basis for generalizing to new voiceless stops.

Our results were mixed. On average, participants generalized to unattested voiceless stops, preferring them to other onsets such as [z]. A closer look at the results revealed that this effect was primarily driven

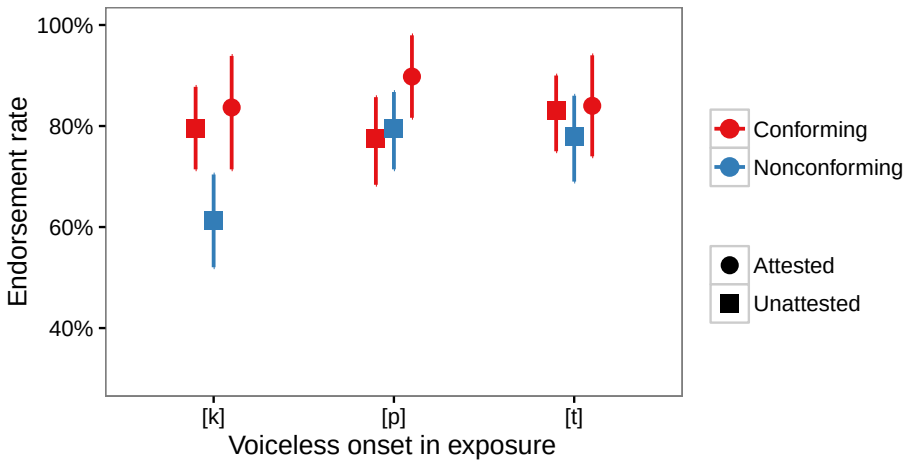

Figure 6: Mean endorsement rates in the Single Type language of Experiment 3, broken down by the voiceless onset in the exposure phase.

by the subset of participants that were exposed to [k] as the voiceless stop; there was no clear evidence for generalization from [p] or [t]. We do not interpret our failure to find statistically significant effects in all subsets of the data as suggesting that participants cannot generalize from a single type; given that the effect size is not very large (in either the Single Type or Two Types language), it would be surprising if all three subsets of the data showed statistically significant differences. At the same time, additional experiments are clearly needed before we can conclude that participants always generalize from a single type of sound.

The Control language was designed to rule out two interpretations of the preference that participants who learned the Single Type language might show for CONF-UNATT over NONCONF-UNATT onsets: first, that participants had a prior preference for voiceless stops, either due to statistical patterns in the English lexicon or for any other reason; and second, that the preference for CONF-UNATT onsets was due to the presence of six APPROX onsets in the exposure phase (though the fact that the two classes of consonants share few phonological features makes this scenario unlikely). After exposure to this language, which included only APPROX onsets, participants did not show a significant difference between the two conditions; if anything, they judged test words with CONF-UNATT onsets as slightly less acceptable than ones with NONCONF-UNATT onsets. This suggests that to the extent that endorsement rates for voiceless stops were higher in the Single Type language, this was due to generalization from the single type of voiceless stop presented in the exposure phase.

In a third language, the Two Types language, the well-formedness of voiceless stop onsets was supported by one token each of two different types of voiceless stops, e.g., both [p] and [k]. Participants again generalized to test words with a CONF-UNATT onset; moreover, they did not distinguish CONF-UNATT from CONF-ATT onsets, replicating the One Set group of Experiments 1 and 2a. There was no significant evidence of a preference for CONF-ATT over CONF-UNATT in the Single Type language either. The two languages differed in that the Two Types language had a single token of each of the two types of voiceless stop, whereas the Single Type language had two tokens of the same type. While this decision served to equalize the number of exposure words across the languages, two tokens of the same onset appear to be sufficient for some onset-specific learning (compare the Two Sets group of Experiments 1 and 2a), which may explain the (nonsignificant) difference in the Single Type language.

In the Control language, which included only six approximant onsets, voiced fricatives were slightly

more likely to be endorsed in the test phase than voiceless stops. This difference, if replicated in more highly powered future experiments, may reflect the fact that voiced fricative and approximants are both voiced consonants. This difference reverses in the Single Type and Two Types languages and becomes a significant preference for voiceless onsets, even though the exposure phase in those language included only two voiceless stop onsets, compared to the six approximant onsets. In other words, participants were more willing to generalize across voiceless stops than from approximants to voiced fricatives. This finding can be interpreted as a preference for generalizing to sounds that differ from the exposure sounds in fewer phonological features, or as a preference for generalizing within natural classes that have fewer members (Albright, 2009) – there are three voiceless stops in English, compared to ten voiced consonants.

6 General discussion

Prior research has shown that speakers can generalize their phonotactic knowledge to novel sounds that share phonological properties with the sounds attested in their language. Similar generalization takes place in artificial language learning experiments: if words in a artificial language often begin with two particular voiceless stops, say [p] and [k], but not with voiced stops, learners will judge novel words that begin with a new voiceless stop (e.g., [t]) as more likely to be words of the language than words that begin with voiced stops. The experiments presented in this paper investigated how generalization to new sounds depends on the amount of exposure to the language that the learner has received.

In Experiments 1 and 2a, participants were divided into four groups that received varying amounts of exposure to an artificial language. In both experiments, participants generalized the phonotactics of the language to words with novel (unattested) consonant patterns, even following brief exposure: they preferred unattested patterns that followed the phonotactic regularities of the language to unattested patterns that did not. By contrast, participants did not start distinguishing the specific sounds they were exposed to from the ones they were not exposed to until they received additional exposure to the language. In other words, participants showed evidence of generalizing (e.g., to new pairs of identical consonants) before they showed evidence of learning any of the specific consonant patterns that supported this generalization (e.g., [p, p]). There was substantial evidence for the "null" hypothesis according to which endorsement rates for attested and unattested consonants were equal following a single exposure set (the Bayes factors were 10.6 in Experiment 1 and 33.2 in Experiment 2a; if the two are combined, the approximate posterior probability of the null hypothesis is 0.97). Finally, generalization to new sounds persisted undiminished despite growing exposure to the language.

In both Experiment 1 and Experiment 2a, the regularity that participants used to generalize to novel sounds was supported by multiple types. In the critical condition of Experiment 3, by contrast, participants were only exposed to a single representative of a phonological class ([p], [t] or [k]). Even when the amount of exposure to the generalization was severely reduced in this way, participants still generalized to sounds that shared phonological properties with that sound, though this effect was only clearly observed for [k].

The rest of the General Discussion addresses the theoretical implications, limitations and potential extensions of these empirical results. Section 6.1 discusses how the results bear on models of phonotactic learning that are based on phonological classes. Section 6.2 discusses alternative interpretations of our results that do not make reference to phonological classes. Section 6.3 clarifies that our experiments do not allow us to delineate all of the precise generalizations that the participants may have entertained. Finally, Section 6.4 addresses the differences and similarities between our results and the results of previous studies of phonotactic generalization.

6.1 Implications for models of probabilistic phonotactics

Three major empirical observations emerge from our experiments. First, participants may be able to generalize from a single type; second, they generalize before they show evidence of distinguishing attested from unattested sounds; and third, they keep generalizing even after a substantial number of exposure sets (up to 16). It is hard to see how the minimal generalization view could be reconciled with an empirical finding of generalization from a single type; however, the current study provides only mixed evidence for the single-type generalization. The implications of the second and third results for computational models are more complicated, and we discuss them in this section. We limit our discussion to models that view phonotactic learning as consisting of the acquisition of a probabilistic model based on phonological features (Adriaans & Kager, 2010; Albright, 2009; Hayes & Wilson, 2008; Linzen & O'Donnell, 2015).

6.1.1 Generalization before sound-specific learning

In the One Set groups of Experiments 1 and 2a, as well as in the Single Type and Two Types languages of Experiment 3, the endorsement rates for the attested sounds and the sounds that participants generalized to were statistically indistinguishable. This is inconsistent with a straightforward implementation of the specific-to-general assumption, in particular in a model like STaGe (Adriaans & Kager, 2010), in which only statistical patterns that are actively used to make phonotactic decisions (word segmentation in the case of STaGe) can give rise to phonotactic generalizations.

Early generalization can be reconciled with the minimal generalization assumption in a model that (1) avoids applying sound-specific patterns to novel words if the number of exposure words that contained that sound was lower than a certain threshold, but (2) uses those sound-specific patterns to form phonotactic generalizations (Albright & Hayes, 2002). If that is the case, knowledge about multiple specific sounds from a class might lead to generalization to that class without a difference in acceptability between attested and unattested sounds (see Linzen & Gallagher, 2014 for simulations).

While non-minimal generalization models predict early generalization, they do not necessarily predict an outcome in which novel sounds that follow the generalization are judged as *equally* well formed as the exposure sounds, as they were in the One Set groups. Under certain assumptions, maximum entropy models predict that attested sounds should always be preferred to unattested ones, regardless of the amount of exposure (for simulations, see Linzen & O'Donnell, 2015). A single exposure to a [b], for example, leads a maximum entropy learner as implemented by Linzen and O'Donnell (2015) to increase some of the weights that apply to other sounds such as [d] (e.g., the weight for voiced stops or for stops); but it will also increase the weights of classes that apply to [b] but not to [d], such as the weight for labials or a weight specific to [b]. Consequently, the attested sound [b] would be preferred to the unattested [d]. The prediction of both a generalization *and* an attestation effect made by the maximum entropy model (as implemented by Linzen and O'Donnell (2015)) is consistent with the empirical endorsement rates after multiple exposure sets, but is inconsistent with the pattern that emerged after minimal exposure.

The absence of an attestation effect after limited exposure may reflect a parsimony bias that encourages the learner to represent the input using fewer phonological classes (Linzen & O'Donnell, 2015; cf. Chomsky & Halle, 1968, p. 337). If the learner has been exposed to five different types of voiced onsets (as in Experiment 1), this bias would lead it to characterize words in the language as beginning with voiced consonants—a single generalization—rather than as beginning with [g], [b], [v], [z] or [ð] (five separate generalizations). As the learner receives more exposure to the language, however, the absence of conforming unattested sounds becomes more apparent, and prompts it to revert to a less parsimonious but more accurate sound-specific representation. Similar sparsity pressures can be incorporated into maximum entropy mod-

els; Hayes and Wilson (2008), for example, implement a feature selection procedure that starts from simpler phonological classes and only adds more complex ones if there is sufficient evidence for them.[5]

At first blush, it may seem that the early acquisition of broader classes could reflect a bias in favor of more general patterns (e.g., identical consonants) and against sound-specific ones (e.g., [k, s]). However, with the exception of the Single Type language in Experiment 3, general and specific patterns never received the same amount of support in our artificial languages. In each exposure set in Experiment 2a, for example, participants heard four words that contained an identical consonant pair, but only one word that contained the specific consonants [k] and [s]. Any advantage of the general patterns, then, could be due to the greater number of examples of those patterns. If anything, participants were able to learn sound-specific patterns from *fewer* examples: for example, in the Two Sets group of Experiment 2a, the endorsement rate for words with specific NONCONF-ATT consonant pairs, which were supported by only two exposure words, was similar to the endorsement rate of identical CONF-UNATT test words, which were supported by eight exposure words (Figure 3).

6.1.2 Sustained generalization

The fact that participants kept generalizing at the same rate even after multiple exposure sets is problematic to models that are sensitive to indirect negative evidence. In Experiment 2a, for example, only four consonants were ever repeated in a word: [p], [ʃ], [g] or [n] (e.g., *papu*). Other consonants in the language, such as [s] or [m], were never repeated. After eight or 16 exposure sets, that absence could be taken to constitute indirect evidence that those consonants cannot be repeated: if they were allowed to be repeated, one would expect them to occasionally be repeated by chance. The Bayesian model of Linzen and O'Donnell (2015) predicts a sharp decline in generalization by the Eight Sets group, in contrast to participants' behavior. Maximum entropy models suffer from a similar problem – in the limit, they are expected to stop generalizing to unattested sounds – although the rate at which they approach this state can depend on various parameters.

The minimal generalization learner (Albright, 2009), on the other hand, does not implement indirect negative evidence: the probability mass reserved to new sounds does not depend on the number of times the attested sounds have been observed. It can therefore capture the sustained generalization pattern.[6] Yet one would expect there to be a limit to speakers' willingness to generalize; English speakers eventually notice the absence of [h]-final words and stop generalizing to those words from words that end with other fricatives such as [f] or [z]. From the empirical perspective, then, it would be useful to determine how robust the sustained generalization pattern is. Would participants continue to generalize even after hundreds of exposure sets? If at some point participants do stop generalizing, that would support probabilistic models that incorporate indirect negative evidence; however, it would still be an important challenge to understand why those models stop generalizing sooner than humans.

[5] It is unclear if the specific procedure advocated by Hayes and Wilson (2008) would be sufficient to simulate the results; we were unable to the run the code available online on our materials since it requires at least 3000 training items. GMECCS, the other published maximum entropy model (Moreton et al., 2015), does not make clear predictions about the relationship between the amount of exposure data and the generalization being acquired; the authors do report gradual convergence towards the target distribution after multiple steps ("trials") of their learning algorithm, but the relationship between the number of trials and the number of observed data points is unclear (hundreds of such "trials" appear to correspond to a single training example). See Linzen and O'Donnell (2015) for an implementation of a maximum entropy model that is sensitive to the amount of training data in a more straightforward fashion.

[6] The adjustment for confidence proposed in Albright and Hayes (2002) and implemented in Linzen and Gallagher (2014) only affects probability estimates in the early stages of acquisition. As mentioned above, it has the opposite effect in the materials of

6.2 Mechanisms of generalization

The empirical pattern across all experiments was unambiguous: participants showed rapid and sustained generalization to words with novel sounds or sound sequences. A range of proposed psychological mechanisms are compatible with this pattern of results, however. We have focused on an interpretation in which participants judged the test words for acceptability by evaluating whether the test words followed one or more probabilistic generalizations extracted during the exposure phase (Albright & Hayes, 2003; Hayes & Wilson, 2008; Frank & Tenenbaum, 2011; Linzen & O'Donnell, 2015; Moreton et al., 2015). Yet the same results may be consistent with a view in which participants evaluate the similarity between the consonant pattern of the test word and their memories of the consonant patterns in the exposure words (Goldinger, 1998; Nosofsky, 1986; Redington & Chater, 1996). Such a similarity metric would need to operate over phonological features rather than pure acoustic similarity (Cristia et al., 2013); to account for the results of Experiment 2a, that similarity metric would also need to make reference to the abstract notion of repetition, to prevent [s, s] from being considered more similar to [s, t] than to [t, t]. Once the representational apparatus is equated between the probabilistic abstraction model and the similarity-based exemplar models, however, the two classes of accounts become difficult to distinguish empirically (Barsalou, 1990; Hahn & Chater, 1998); indeed, exemplar models have been interpreted as a process-level implementation of the probabilistic abstraction approach (Shi, Griffiths, Feldman, & Sanborn, 2010). We therefore hesitate to interpret our results as providing support for either mechanistic characterization of generalization.

Did participants generalize using independently represented phonotactic patterns (either rule-based or exemplar-based), or did they use analogy to whole exposure words, matching the test words to their (possibly inaccurate) memories of particular exposure words (Bailey & Hahn, 2001; White, Yee, Blumstein, & Morgan, 2013)? Since our test words were all novel – even CONF-ATT test words differed from all exposure words at least in their vowel patterns – our paradigm does not allow us to probe participants' memory of particular exposure words. We believe, however, is that it is unlikely that participants remembered a significant fraction of the exposure words. Words were never repeated more than once in exposure; the high variability of the vowel patterns (and therefore the particular words) is likely to have encouraged learning of the consonant patterns rather than learning of particular words (Gómez, 2002). Indeed, although it is probably more difficult to remember 64 different words (in the Eight Sets group) than eight different words (in the One Set group), participants in the Eight Sets group showed better learning outcomes than those in the One Set group. It is likely that participants were not particularly motivated to memorize individual words: those words were not paired with a meaning, and the instructions emphasized that the test phase would consist entirely of novel words. Finally, similarity to particular exposure words is an unlikely explanation for the results of Experiment 2a, where participants generalized to CONF-UNATT words that did not share a single sound with the exposure words (e.g., from *pipa* to *keku*). These considerations aside, we acknowledge that the role of memory for particular exposure words is an understudied problem in phonotactic learning experiments; future experiments manipulating the factors mentioned above may be able to distinguish lexicon-based generalization from independently represented phonotactic knowledge.

6.3 The extent of generalization

All of the experiments reported in this paper followed the same logic: they tested whether participants preferred novel sounds from a phonological class that contained the exposure sounds to novel sounds outside

Experiment 2a: it boosts the probability of generalization and reduces the probability of attested items.

that class. At the same time, the results should not be interpreted as indicating that participants extracted a *single* phonological pattern from the exposure sounds. For instance, while the results of Experiment 1 indicate only that after exposure to [k t f θ p s] learners generalized to other voiceless obstruents (the minimal class that included all exposure sounds), they do not provide evidence that participants restricted their generalization *only* to onsets that belonged to that class. Indeed, it is plausible that participants would also have generalized to classes that only include some of the exposure sounds, such as dorsal stops (a class that includes the exposure onset [k], but also [g] and others) or fricatives (a class that includes [f], as well as [v] and others).

The single-class interpretation is even less applicable to the other experiments: in Experiment 2a, participants generalized the consonant repetition pattern to new sounds even though that pattern only held of half of the exposure words. In Experiment 3, participants were only exposed to a single type of voiceless stop (e.g., [k]); there were clearly not in a position to guess the dimension along which they would be expected to generalize in the test phase (voicing, place of articulation, manner...). Rather, it is likely that participants considered multiple probabilistic phonotactic patterns that were compatible with some or all of the exposure items; in the case of [k] in Experiment 3, those patterns may have included voiceless stops, dorsal stops, dorsal consonants and so on.

The consistently high endorsement rates for NONCONF-UNATT test words — items that did not belong to the narrowest phonological class supported by the exposure words, but nevertheless shared many properties with them — can also be taken to suggest that participants generalized to those words as well, though to a lesser extent than to CONF-UNATT test words. In the future, concrete evidence for graded generalization could be obtained by comparing three or more classes of unattested sounds that are increasingly different from the exposure items; for example, if the exposure sounds were voiced stops, the test conditions might be other voiced stops, voiceless stops, non-stop consonants (e.g., fricatives), and finally vowels.

Finally, in a given exposure group of Experiments 1 and 2a all patterns were represented by the exact same number of exemplars: a participant in the Four Exposures group of Experiment 1, for example, heard exactly four words starting with each of the five onsets. This uniform distribution over attested types may have made the generalization particularly salient, leading to faster and more sustained generalization than would be the case if the distribution was not uniform (for instance, Zipfian); this hypothesis can be tested in future work.

6.4 Previous studies of phonotactic generalization

Our finding that participants generalized to new sounds is in line with the results of several other studies (Cristia et al., 2013; Finley & Badecker, 2009; Finley, 2011; Gallagher, 2013). However, those studies tested participants after extensive exposure to the language: 160 words (Cristia et al., 2013), 212 words (Gallagher, 2013) or 120 words (Finley & Badecker, 2009; Finley, 2011). Our study enriches the empirical picture by charting how the generalizations that participants make depend on the amount of exposure to the artificial language, in particular when given a very small amount of exposure: participants in the One Set condition in Experiment 1 received only five exposure words, a fraction of the number of exposure words used in previous studies.

Some learning experiments that used different paradigms from ours have found that participants did not generalize as readily as in our experiments. In a study that assessed phonotactic learning using speech errors in production, participants only generalized phonotactic constraints to new sounds if a period of sleep intervened between the exposure and test sessions (Gaskell et al., 2014). Two studies that examined phonotactic learning in the context of a morphological alternation also did not report generalization to new segments (Peperkamp et al., 2006; Peperkamp & Dupoux, 2007).

Phonotactic learning experiments vary along more subtle methodological dimensions as well. We asked our participants whether they believed that the test words could be part of the language that they had learned. This task is similar to the wordlikeness task used to investigate natural language phonotactics (Coleman & Pierrehumbert, 1997; Bailey & Hahn, 2001) and to the tasks used in learning artificial grammars of word sequences (among many others, Gomez, 1997). Some phonotactic learning experiments have used different tasks. Cristia et al. (2013), for example, asked their participants how frequently they had heard the test items in the exposure phase; even though all test items were novel, participants provided different familiarity judgments to test items of different conditions. It is quite possible that participants generalize more conservatively when judging a test item for familiarity than when judging it for acceptability. In the future, it would be useful to perform a direct comparison across tasks with the same language and training regime.

All of our participants were English-speaking adults. As such, our experiments can be argued to be a closer approximation of second language learning than of first language acquisition. At the same time, we are encouraged by the fact that our findings converge with the results of infant studies. Six month old infants exposed to a language very similar to the one used in Experiment 1 showed a similar behavior to the adult participants in the One Set group of Experiments 1 and 2a: they looked longer at words that started with CONF-UNATT than NONCONF-UNATT onsets, but did not distinguish CONF-UNATT from CONF-ATT onsets (Cristia & Peperkamp, 2012). The infants in that study were exposed to a much larger number of exposure words than the adults in our One Set condition (54 as opposed to five), making it difficult to know how rapidly they generalized. Stronger evidence for rapid phonotactic generalization in infants was obtained in two recent experiments by Gerken and colleagues. Nine-month-olds who have been exposed to a single word with a duplicated syllable (*leledi*), repeated a few times, preferred novel words with a similar structure, suggesting that they learned a reduplication rule from a single example (Gerken, Dawson, Chatila, & Tenenbaum, 2015); this is consistent with the finding of single-type generalization in Experiment 3. A second study showed that 11-month-olds were able to extract a generalization from only four words (which represented different types), in line with the adults in the One Set condition of Experiment 1 (Gerken & Knight, 2015).

7 Conclusion

This paper reported on a series of artificial language experiments that investigated the time course of phonotactic generalization. The experiments showed that participants can generalize beyond the specific sounds that occurred in the language following a very short exposure session; in fact, they generalized before they showed evidence of recognizing individual exposure sounds. This was the case regardless of whether the phonotactic regularity that was generalized to new sounds was categorical or probabilistic, and of whether it was based on a phonological class or an identity relation across segments. Generalization continued undiminished despite growing exposure to the language. Finally, there was some evidence that participants can generalize to new sounds based on a single type of sound only; single-type generalization may be more likely with some exposure sounds than others.

Our results are not fully consistent with any of the existing models of phonotactics: rapid generalization given limited exposure to the language is inconsistent with minimal generalization models (Adriaans & Kager, 2010; Albright, 2009), and the finding of sustained generalization after additional exposure is inconsistent with models that make strong use of indirect negative evidence (Hayes & Wilson, 2008; Linzen & O'Donnell, 2015; Moreton et al., 2015). Our findings can therefore inform the development of more adequate models of phonotactics. More generally, we suggest that models of phonotactics should make

explicit predictions concerning the relationship between the amount of training data and the generalizations extracted from the data.

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*(3), 311–331.

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

Albright, A., & Hayes, B. (2002). Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (pp. 58–69). Stroudsburg, PA: Association for Computational Linguistics.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 3, pp. 61–88). Hillsdale, NJ: Erlbaum.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Becker, M., & Levine, J. (2010). Experigen: An online experiment platform. *Available (April 2013) at https://github. com/tlozoot/experigen*.

Berent, I., Marcus, G., Shimron, J., & Gafos, A. (2002). The scope of linguistic generalizations: Evidence from Hebrew word formation. *Cognition*, *83*(2), 113–139.

Berwick, R. C. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

Chambers, K. E., Onishi, K. H., Fisher, C., et al. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *arXiv preprint cmp-lg/9707017*.

Cristia, A., Mielke, J., Daland, R., & Peperkamp, S. (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, *4*(2), 259–285.

Cristia, A., & Peperkamp, S. (2012). Generalizing without encoding specifics: Infants infer phonotactic patterns on sound classes. In A. K. Biller, E. Y. Chung, & A. E. Kimball (Eds.), *Proceedings of the 36th Annual Boston University Conference on Language Development (BUCLD 36)* (pp. 126–138). Somerville, MA: Cascadilla Press.

Cristià, A., & Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development*, *4*(3), 203–227.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*(2), 197–234.

Dell, F. (1981). On the learnability of optional phonological rules. *Linguistic Inquiry*, *12*(1), 31–37.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Finley, S. (2011). Generalization to novel consonants in artificial grammar learning. In C. H. Laura Carlson & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 318–23).

Finley, S., & Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, *61*(3), 423–437.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371.

Friederici, A. D., & Wessels, J. M. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics*, *54*(3), 287–295.

Gallagher, G. (2013). Learning the identity effect as an artificial language: Bias and generalisation. *Phonology*, *30*(2), 253–295.

Gaskell, M. G., Warker, J., Lindsay, S., Frost, R., Guest, J., Snowdon, R., & Stackhouse, A. (2014). Sleep underpins the plasticity of language production. *Psychological Science*, 1457–1465.

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74.

Gerken, L., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, *4*(3), 228–248.

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science*, *18*(1), 80–89.

Gerken, L., & Knight, S. (2015). Infants generalize from just (the right) four words. *Cognition*, *143*, 187–192.

Gervain, J. (2014). Early rule-learning ability and language acquisition. In F. Lowenthal & L. Lefebvre (Eds.), *Language and recursion* (pp. 89–99). New York: Springer.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Gomez, R. L. (1997). Transfer and complexity in artificial grammar learning. *Cognitive Psychology*, *33*(2), 154–207.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Hahn, U., & Chater, N. (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, *65*(2), 197–230.

Hale, M., & Reiss, C. (2003). The Subset Principle in phonology: Why the tabula can't be rasa. *Journal of Linguistics*, *39*(2), 219–244.

Hayes, B. (2011). *Introductory phonology*. Malden, MA and Oxford: Wiley-Blackwell.

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jessen, M., & Ringen, C. (2002). Laryngeal features in german. *Phonology*, *19*(2), 189–218.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*(3), 402–420.

Jusczyk, P. W., & Luce, P. A. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645.

Kapatsinski, V. (2014). What is grammar like? A usage-based constructionist perspective. *LiLT (Linguistic Issues in Language Technology)*, *11*.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Levy, R. (2014). Using R formulae to test for main effects in the presence of higher-order interactions. *arXiv preprint arXiv:1405.2094*.

Linzen, T., & Gallagher, G. (2014). The timecourse of generalization in phonotactic learning. In J. Kingston, C. Moore-Cantwell, J. Pater, & R. Staub (Eds.), *Proceedings of Phonology 2013.* Washington, DC: Linguistic Society of America.

Linzen, T., & O'Donnell, T. J. (2015). A model of rapid phonotactic generalization. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP) 2015* (pp. 1126–1131).

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

McMahon, A. M. (2002). *An introduction to English phonology*. Edinburgh: Edinburgh University Press.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*(1), 21–46.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*(1), 83–127.

Moreton, E. (2012). Inter- and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*, 165–183.

Moreton, E., Pater, J., & Pertsova, K. (2015). Phonological concept learning. *Cognitive Science*, 1–66.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Peperkamp, S., & Dupoux, E. (2007). Learning the mapping from surface to underlying representations in an artificial language. *Laboratory Phonology*, *9*, 315–338.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*(3), B31–B41.

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*(2), 123–138.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30–54.

Richtsmeier, P. T. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology*, *2*(1), 157–183.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494.

Scholes, R. J. (1966). *Phonotactic grammaticality*. The Hague: Mouton.

Seidl, A., & Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development*, *1*(3-4), 289–316.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464.

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Vitevitch, M., Luce, P., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, *40*(1), 47–62.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 387–398.

White, K. S., Yee, E., Blumstein, S. E., & Morgan, J. L. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, *68*(4), 362–378.

Wilson, C. (2003). Experimental investigation of phonological naturalness. In G. Garding & M. Tsujimura (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics* (pp. 533–546).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.