Rapid generalization in phonotactic learning

Tal Linzen and Gillian Gallagher

Department of Linguistics, New York University

August 24, 2015

Abstract

The phonotactics of a language concerns the well-formedness of strings of sounds as potential words (e.g., *paim* is a better potential word of English than *mlpemr*). Speakers' phonotactic judgments are informed not only by the distribution of particular sounds ([b] or [g]) but also by the distribution of classes of sounds (e.g., voiced stops). In a series of artificial language experiments, we investigate how such generalizations over classes of sounds are acquired, focusing on evaluating the proposal that generalizations must be acquired in a specific-to-general sequence – i.e., that learners must first learn the statistics of multiple individual sounds that belong to a class before they can generalize to the class. Contrary to this proposal, learners acquired knowledge over classes earlier than sound-specific knowledge, and showed an ability to generalize to a class based on a single example of the class. We discuss the implications of our findings for computational models of phonotactic learning.

1 Introduction[1]

Natural languages often place restrictions on how sounds can combine to form words. The velar nasal [ŋ], for example, can occur in the coda of an English syllable, as in *ring* [ɹɪŋ] or *finger* [fɪŋgəɹ], but not in its onset:[2] Words like **ngir* [ŋɪɹ] do not exist in English (McMahon, 2002). English speakers

[1]Experiments 2a and 2b were published in the Proceedings of Phonology 2013 in a different form (Linzen & Gallagher, 2014). We thanks Frans Adriaans, Michael C. Frank, Timothy O'Donnell and Todd Gureckis for discussion, as well as audiences at the 8th Northeast Computational Phonology Circle and the 36th Annual Cognitive Science Society Meeting.

[2]The onset of a syllable consists of the consonant or consonants that precede its nucleus (the nucleus is generally a vowel); its coda consists of the consonant(s) that follow the nucleus. In the syllable [flɪp], for example, [fl] is the onset, [ɪ] is the nucleus and [p] is the coda. We use an asterisk to represent phonotactically illegal words.

do not typically consider this to be an accidental gap, and judge words that start with a [ŋ] 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 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 that share articulatory or perceptual features tend to have similar phonotactic distributions. German, for example, allows all voiced consonants (e.g., [b] or [g]) to occur anywhere in the word except its end: bal] is a valid German word, but *[*lab*] isn't. Although such patterns can in principle be captured by a disjunction of multiple sound-specific patterns, learners represent them as class-wide patterns as well (Saffran & Thiessen, 2003; Cristià & Seidl, 2008). For example, English speakers rate *srip* as a better potential word of English than *mbip* (Scholes, 1966; Daland et al., 2011), even though neither of the sequences [sɹ] and [mb] occur at the onset of English words. A plausible interpretation of this finding is that speakers generalize from the attested English onsets [sl] and [ʃɹ] (e.g., *slip* or *shrewd*) to the wider class of strident-liquid onsets, which also includes [sɹ]. On the other hand, there are no attested sonorant-stop onset sequences that would give rise to an analogous generalization that would apply to [mb] (Albright, 2009).

1.1 The time course of generalization

In line with these observations, models of phonotactic knowledge that incorporate natural-class based generalizations typically predict behavioral results better than models that do not (Adriaans & Kager, 2010; Albright, 2009; Berent, Wilson, Marcus, & Bemis, 2012; Hayes & Wilson, 2008). The relevant behavioral results are generally obtained by presenting novel strings to speakers of a given language and asking them for implicit or explicit judgments of how likely those words are to be words of that language; in other words, models are tested on their ability to simulate the end-state of language learning, after considerable exposure to the language. Conversely, there is little empirical data on the *time course* of phonotactic learning, even though those models make different assumptions about that time course. The goal of this paper is to characterize the process by which phonotactic generalizations are learned, with an eye to constraining models of phonotactics.

We contrast two views of the order of acquisition of item-specific knowledge and broader generalizations (see also Cristia & Peperkamp, 2012; Kapatsinski, 2014). In one view, which we will term *specific-to-general learning*, learners must first acquire knowledge about specific segments. Once they have noticed the commonalities among specific segments, they can form a generalization that may encompass unattested phonotactic structure in addition to the attested ones that gave rise to the generalization. For example, when acquiring the phonotactics of English, learners must first learn that English syllables can start with [b] and that they can start with [g] before they can make the generalization that English syllables can start with a voiced stop (Albright & Hayes, 2003; Albright, 2009; Adriaans & Kager, 2010).

StaGe (Adriaans & Kager, 2010) is a particularly clear example of the specific-to-general view. It includes two distinct modules: a statistical learning module and a generalization module (for a related proposal, see Thiessen, Kronstein, & Hufnagle, 2013). The statistical learning module tracks the frequency of two-sound word-initial sequences (e.g., [bl] or [mp]).[3] If the frequency of the sequence is significantly lower than expected from the frequencies of the sounds that make up the sequence (typically less than half of the expected frequency), the model induces a constraint against having the sequence as a word-initial cluster. Whenever the statistical learning module has acquired constraints against two sound sequences that differ in exactly one phonological feature, the generalization module constructs a constraint that abstract away from that feature. For example, if the statistical learning module has acquired a constraint against [bl] and [gl], which differ only in the place of articulation of the first consonant, the generalization module will construct a constraint against all voiced stops followed by an [l] (which can then apply to [dl], for instance).

A second view, which we will term *simultaneous learning*, does not assume that the learning of a generalization presupposes the existence of sound-specific knowledge. The existence of a [b]-initial syllable in the input supports both a segment-specific pattern (e.g., syllables can start with [b]) and a more general one (syllables can start with a voiced stop, or even a stop in general). Consequently, learners may acquire class-wide patterns before any of their segment-specific instances. This view underlies Maximum Entropy models (Hayes & Wilson, 2008; Pater & Moreton, 2012) and PAIM (Linzen & O'Donnell, 2015).

1.2 Overview of the paper

This paper reports the results of a series of artificial language learning experiments that probe the time course of phonotactic generalization. In Experiment 1, participants were taught a language in which all word onsets had the same value for the voicing feature (e.g., all were voiced). This knowledge could

[3] StaGe is a model of phonotactics-based word segmentation. We abstract away from the details related to word segmentation and focus on the phonotactic learning aspect of the model.

be represented either as a disjunction of segment-specific patterns (onsets can be [b], [d] or [g]) or as an abstract pattern (onsets are voiced). Participants were divided into several groups, each of which received a different amount of exposure to the language. After the exposure phase, participants judged novel test words for acceptability. If participants preferred test words whose onset appeared in exposure to test words with a new onset, that would constitute evidence for learning of the specific phonotactic patterns. Conversely, a preference for test words whose onset shared the voicing value with the exposure words to test words that didn't, regardless of whether the specific onset appeared in the exposure phase, would provide evidence for learning of the abstract pattern. Participants showed evidence of learning the abstract pattern before they showed evidence of learning the specific ones; the results of Experiment 1 were therefore inconsistent with the specific-to-general view.

The abstract regularity in Experiment 1 was a categorical phonotactic restriction based on a phonetic feature. Experiment 2a tested the generality of the findings by teaching participants a language with an abstract generalization that is not tied to a phonetic feature—specifically, identity between two consonants. Moreover, this regularity was probabilistic rather than categorical. The results were qualitatively similar to the results of Experiment 1. Experiment 2b taught participants a control language whose goal was to verify that the results of Experiment 2a were indeed due to learning rather than pre-existing biases.

Finally, Experiment 3 tested whether learners can acquire a general phonotactic regularity based on a single instance of the regularity. Participants indeed generalized to sounds that were not in the exposure set but shared a phonetic feature with the single attested exposure sound, indicating that generalization to an abstract pattern does not require extracting commonalities between multiple specific patterns.

2 Experiment 1: A natural-class based generalization

The artificial language used in this experiment was modeled after the one used by Cristia, Mielke, Daland, and Peperkamp (2013). It had a categorical natural-class based phonotactic generalization: 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 on words of three types:

1. Conforming attested onset (CONF-ATT): words whose onset appeared as the onset of one of the exposure words. Since the phonotactic pattern was categorical, all such onsets conformed to the generalization.
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 test words were distinct from the exposure words. This was the case even for CONF-ATT test words, where the onset (C_1) was shared with some of the exposure words, but the full word ($C_1V_1C_2V_2$) 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. Participants were not given any indication that the exposure words were organized into sets. A detailed description is given in the Materials section below; see Table 1 for examples.

The focus of the study is the relationship between the amount of exposure to the language that participants receive and the knowledge that they extract from that input. The specific-to-general view would be most clearly supported if at first participants only distinguished CONF-ATT onsets from all other onsets after a small amount of exposure, and then began generalizing to CONF-UNATT onsets as they received more exposure. The simultaneous view would be most strongly supported if participants first acquired the voicing generalization and only then the specific onsets; that is, if they showed evidence of distinguishing CONF-ATT and CONF-UNATT onsets on the one hand from NONCONF-UNATT onsets on the other hand before they showed evidence of recognizing CONF-ATT onsets. Finally, an outcome where participants started making both of these distinctions at the same time would be compatible with both views.

2.1 Method

2.1.1 Materials and procedure

The materials were a simplified version of those used by Cristia et al. (2013), adapted to English-speaking participants.[4] The onsets of all of the stimuli used in the experiment were drawn from the set of six voiced

[4]Specifically, the fricatives [ʒ] and [ʃ] were replaced with [ð] and [θ], respectively, since [ʒ] is rare in onset position in English. Furthermore, Cristia et al. (2013) had two types of nonconforming test onsets, "near" and "far"; our NONCONF-UNATT test condition

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

(a)

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

ATT (attested): onset consonant (but not the full word) was encountered in exposure phase.
UNATT (unattested): onset consonant were not encountered in exposure phase.
CONF (conforming): onset consonant conforms to the abstract pattern.
NONCONF (nonconforming): onset consonant does not conform to the abstract pattern.

obstruents [b], [d], [g], [ð], [v] and [z] or from the set of six voiceless onsets [p], [t], [k], [θ], [f] and [s]. Words of the format $C_1V_1C_2V_2$ were created with all possible combination of these onsets as 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 in the language, as in all others languages used in this paper, were stressed on their first syllable. The words 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 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 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).

only included their "far" onsets.

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 (www.mturk.com), a crowdsourcing site that enables recruiting a large number of participants for a modest cost. Results obtained using Mechanical Turk have been repeatedly shown to replicate established findings from the experimental behavioral research literature (Crump, McDonnell, & Gureckis, 2013); this inspires confidence in the platform's adequacy as a source of participants for behavioral experiments. 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 in a short demographic survey at the end of the experiment what their native language was; 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. During the exposure phase, the participants listened to words from the artificial language. 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 sound 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).

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 1.1.6 of the lme4 package in R (Bates, Maechler, & Bolker, 2014). To simplify the exposition, we present two separate sets of analyses, one comparing

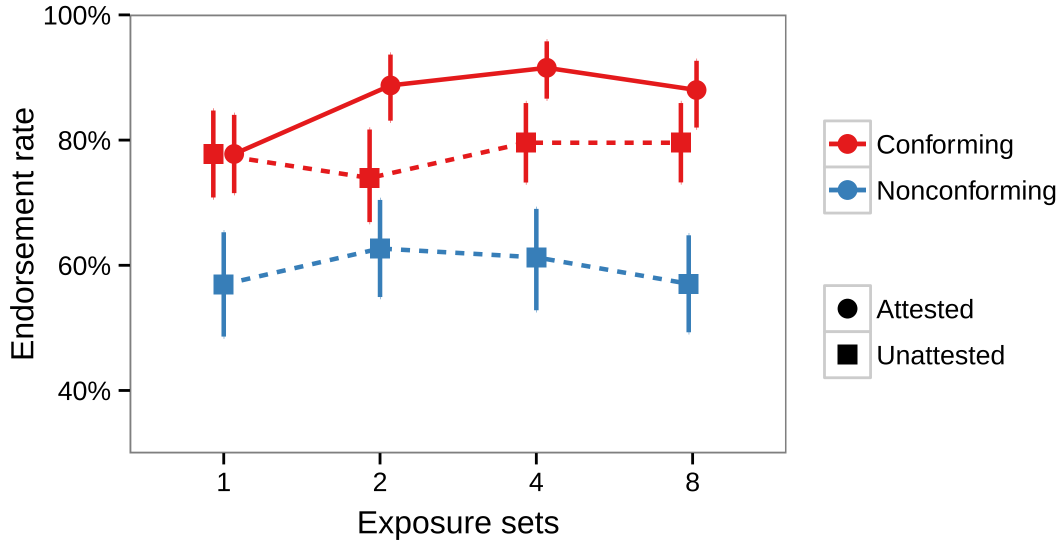

Figure 1: Mean endorsement rates for Experiment 1. Error bars represent bootstrapped 95% confidence intervals.

NONCONF-UNATT to CONF-UNATT onsets and another comparing CONF-UNATT to CONF-ATT onsets. We fitted two types of models: full models, which included all participants, and within-group models, which only included participants in a given group (e.g., the Two Sets group). Fixed effects in the full models included the group as a four-level factor and the onset type as a two-level factor. 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. Statistical significance was assessed using the log-likelihood ratio test: Terms were added sequentially to the model and the improvement in log-likelihood was assessed using the chi-squared distribution (this is the LMEM equivalent of a Type I ANOVA).

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. Endorsement rates were generally high: Even words that started with a NONCONF-UNATT onset were rated as acceptable around 60% of the time. This likely reflects the fact that apart from the onset these words matched the exposure words in all other respects; for example, word length, syllable structure, vowel pattern and identity of medial consonant were all consistent with the exposure words. Thus, all of the testing words received some support from the exposure data. We predict that words with novel syllable structures (e.g., *kes*) or longer words (e.g., *lumpumilu*) will be endorsed at a lower rate.

2.3.1 CONF-ATT vs. CONF-UNATT

In the full model, which included all participants, there was a significant main effect of onset type ($\chi^2(1) = 25.76$, $p < .001$), such that CONF-ATT onsets were rated more highly than CONF-UNATT ones. The main effect of group did not reach significance ($\chi^2(3) = 4.85$, $p = .18$), indicating that the average rating collapsing across these two onset types was similar across exposure groups. The interaction between onset type and group was significant ($\chi^2(3) = 8.79, p = .03$). Within-group models showed that this interaction was driven by the absence of a significant preference for CONF-ATT onsets in the One Set group ($\chi^2(1) = .95$, $p = .33$), compared with a significant preference for CONF-ATT onsets in the Two and Eight Sets groups and a marginal 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: $\chi^2(1) = 17.8, p < .001$).

2.3.2 NONCONF-UNATT vs. CONF-UNATT

CONF-UNATT words were rated more highly than NONCONF-UNATT ones ($\chi^2(1) = 32.27$, $p < .001$). The main effect of group was not significant ($\chi^2(3) = 0.57$, $p = .9$) and neither was the interaction between group and type ($\chi^2(3) = 2.82, p = .42$). 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 only marginally significant (One Set: $\chi^2(1) = 12, p < .001$; Two Sets: $\chi^2(1) = 2.7, p = .1$; Four Sets: $\chi^2(1) = 8.08, p = .004$; Eight Sets: $\chi^2(1) = 11.2, p < .001$). Since the interaction was not significant, we do not interpret the difference between the Two Sets group and the other groups any further.

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 onsets of exposure words to onsets with the opposite voicing. Conversely, participants did not start distinguishing CONF-ATT from CONF-UNATT onsets until after two or more exposures. This pattern of results favors the simultaneous view over the specific-to-general view: Participants in the One Set group, who did not show evidence of learning any of the individual onsets, still preferred onsets conforming to the generalization to onsets that did not.

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 evidence that not all generalization-conforming onsets are possible in the language, then, participants continued to generalize beyond the attested onsets.

3 Experiment 2a: A probabilistic abstract generalization

Participants in Experiment 1 showed evidence of learning a broad regularity (voiced stops can be onsets) before showing evidence of learning narrow regularities (e.g., [b] can be an onset). The broad regularity in Experiment 1 had two properties that may limit the generality of the conclusions that can be drawn from these results.

First, the generalization in Experiment 1 was categorical: It held of all of the words in the language. 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: They also keep track of the relative frequencies of the possible sounds and sound sequences. Neither of the nonwords *riss* [ɹɪs] and *yowdge* [jaʊdʒ], for example, contains any sounds or sound sequences that are unattested in English; yet *riss*, which is comprised of frequent sound sequences, is judged to be a more likely potential word of English than *yowdge* (Coleman & Pierrehumbert, 1997).

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 (e.g., "voiced stops are illegal codas" in German), some generalizations involve *relations* across sounds. A prominent example is the constraint against identical root-initial consonants in Semitic languages (Greenberg, 1950; McCarthy, 1986; Frisch & Zawaydeh, 2001; Rose & King, 2007). In Hebrew, as in other Semitic languages, typical roots consist of three consonants (e.g., *ktb* 'write'). While roots with two identical consonants are in general fairly common (*smm* 'drug', *zrz* 'hasten'), cases in which the two identical consonants are the first two consonants of the root are rare to nonexistent (**ssm, *zzr*). Hebrew speakers are sensitive to this restriction. In a lexical decision task, for example, novel roots with initial duplication (e.g., *ssk*) are identified as nonwords faster than are novel roots with final duplication (*kss*); the preference for *kss* over *ssk* is stronger than the phonotactic probabilities of the sound sequences that make up those roots would predict (Berent, Shimron, & Vaknin, 2001). More strikingly, root-initial duplicated consonants are disfavored even when they are made up of consonants that do not exist in Hebrew. Although [w] is not part of the sound inventory of Hebrew, and consequently the sequence *ww* has a phonotactic probability of 0 both root-finally and root-initially, Hebrew speakers recognize **wwp* as a nonword faster than they recognize **pww* (Berent, Marcus, Shimron, & Gafos, 2002). This suggests that Hebrew speakers generalize over individual sequences of duplicated consonants to form a constraint disfavoring any root-initial duplicated consonant: *XXY, for any X and Y. Relational phonotactic generalizations have been documented in such diverse languages as Yucatec Mayan, Muna (an Austonesian language) and Peruvian Aymara (for a recent review, see Gallagher, 2013).

To replicate the findings of Experiment 1 and broaden the scope of the conclusions that can be drawn from those findings, 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 narrow 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 identical consonants, e.g., C_1 = [p] and C_2 = [p] (as in *pepu*), or C_1 = [s] and C_2 = [s] (as in *sase*).

While the phonotactics of the language can be captured precisely using these eight narrow regularities, it was also the case that half of the words in the language conformed to 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 in Experiment 1, exposure sets were created that included exactly one word that conformed to 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 presented with new words that had either consonant pairs that were familiar from the exposure phase (ATT) or new consonant pairs (UNATT), and asked them to judge whether the testing words could belong to the language they had learned. Half of the new consonant pairs in testing had identical consonants (CONF) and half had non-identical consonants (NONCONF). By contrast with Experiment 1, it was possible to construct ATTESTED-NONCONFORMING test words, since the generalization was probabilistic and held only of half of the exposure words. This led a fully crossed design that allowed us to test for the independent contribution of the broad and narrow regularities.

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

Table 2: Materials presented to the participants in Experiment 2a. The table shows a complete exposure and test set for the One Set group.

If participants learned the broad regularity, namely that identical consonant pairs are particularly common in the language, they should prefer words with identical consonant pairs to words with non-identical consonants. If participants learned the narrow consonant-specific regularities, they should prefer words with attested consonant pairs to words with unattested ones.

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 and eight with new consonant pairs. Each of the individual consonants C_1 and C_2 in the new 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.

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. There were 16 test words in all groups, one item with each consonant pair. As in Experiment 1, the specific words from 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.

Items were pseudo-randomized in blocks as in Experiment 1.

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 were unable to include a by-item random effect due to model convergence issues. As before, p-values were calculated using the chi-square approximation to likelihood ratio tests in a stepwise regression. In the within-group models as well it was necessary to simplify the random effect structure. All inferences involved models with random intercepts for subjects and for consonant pair. For inferences involving Attestation, we only included a random slope for the factor; likewise, for inferences involving Conformity, we only included a random slope for Conformity. The interaction between Attestation and Nonconformity was assessed in the model that only had an Attestation random slope, again for convergence issues. Inferences without a random slope can be anti-conservative (Barr, Levy, Scheepers, & Tily, 2013). As such, we can trust the model when it finds that an interaction is nonsignificant, but cannot necessarily trust it when it finds the interaction to be significant.

3.2 Results

3.2.1 Full model

Figure 2 illustrates the mean endorsement rates for each group and condition. The full statistical model yielded a main effect of group ($\chi^2(3) = 45.12$, $p < .001$), reflecting the fact that endorsement rates were higher for participants who received more exposure to the language. There was also a main effect of Attestation, reflecting higher average endorsement rates for words with ATT than for words with UNATT consonants ($\chi^2(1) = 34.04$, $p < .001$), and a main effect of Conformity, reflecting higher average endorsement rates for CONF than for NONCONF words ($\chi^2(1) = 35.46$, $p < .001$).

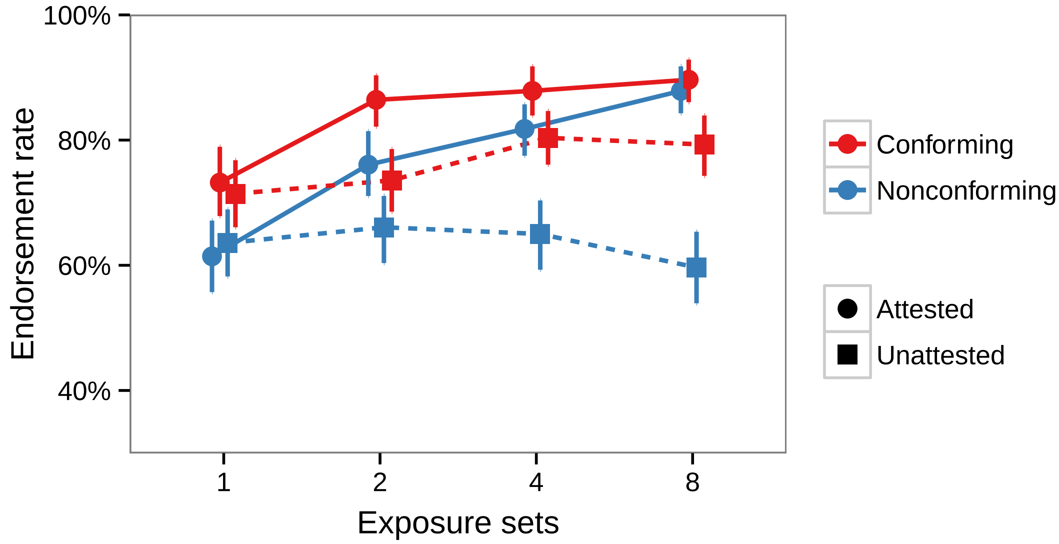

Figure 2: Mean endorsement rates for Experiment 2a. Error bars represent bootstrapped 95% confidence intervals.

The main effect of Attestation was modulated by an interaction with group ($\chi^2(3) = 37.9, p < .001$), 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 ($\chi^2(3) = 1.12$, $p = .77$), and neither was the interaction between Conformity and Attestation ($\chi^2(1) = 0$, $p = 1$). The interpretation of these findings is complicated by the significant three-way interaction ($\chi^2(3) = 8.59$, $p = .03$); Figure 2 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 remained 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 rated significantly higher than NONCONF ones ($\chi^2(1) = 5.76$, $p = .02$). The main effect of Attestation and the interaction did not reach significance (Attestation: $\chi^2(1) = 0$, $p = .98$; interaction: $\chi^2(1) = .28$, $p = .6$), suggesting that the narrow phonotactic regularities did not affect endorsement rates. There was no sign of a nonsignificant numerical trend towards an effect of Attestation (endorsement rates were: CONF-ATT: 73%; CONF-UNATT: 71%; NONCONF-ATT: 61%; NONCONF-UNATT: 64%).

Two Sets: Test words with ATT consonants were rated as acceptable significantly more often than ones with UNATT consonants ($\chi^2(1) = 12.1$, $p < .001$). The effect of Conformity was marginally significant ($\chi^2(1) = 3.56$, $p = .06$), and the interaction was nonsignificant ($\chi^2(1) = .99$, $p = .32$).

Four Sets: Both of the main effects reached significance; the interaction was again nonsignificant (Conformity: $\chi^2(1) = 4.97$, $p = .03$; Attestation: $\chi^2(1) = 11.49$, $p < .001$).

Eight Sets: The effect of Attestation was highly significant ($\chi^2(1) = 20.23$, $p < .001$). Conformity had a marginally significant effect ($\chi^2(1) = 3.49$, $p = .06$); however, there was a significant interaction between Attestation and Conformity ($\chi^2(1) = 4.23$, $p = .04$). As mentioned above, the absence of a random slope for the interaction makes the interpretation of this p-value problematic. However, separate models fitted within ATT and UNATT items (both with random found that Conformity had a highly significant effect for UNATT items ($\chi^2(1) = 17.6$, $p < .001$) but no discernible effect for ATT ones ($\chi^2(1) = 0.34$, $p = .56$).

3.3 Discussion

Participants showed evidence of learning the broad regularity over identical consonant pairs before they showed evidence of learning the narrower regularities over individual C_1–C_2 pairs. 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 this pair of identical consonants was presented in the exposure phase. Participants did not start showing evidence of learning individual consonant pairs until they received at least two sets of exposure. As in Experiment 1, this pattern of results suggests that broad regularities can be learned before narrower instances of those regularities.

Also echoing 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 other groups (One, Two and Four Sets); this suggests that the dip in endorsement rates for NONCONF-UNATT in the Eight Sets group visible in Figure 2 was spurious (CONF-ATT: 92%; CONF-UNATT: 79%; NONCONF-ATT: 89%; NONCONF-UNATT: 67%). Only the main effect of Attestetedness was significant ($\chi^2(1) = 28.72$, $p < .001$); the main effect of Conformity and the interactions were not (Conformity: $\chi^2(1) = 1.13$, $p = .29$; interaction: $\chi^2(1) = 1.23$, $p = .27$). The simple effect of Conformity was significant within UNATT words ($\chi^2(1) = 4.77$, $p = 0.03$) but not within ATT ones ($\chi^2(1) = .05$, $p = .83$). In sum, statistical evidence for generalization to CONF-UNATT words remained even for participants who received 16 exposure sets; the fact that this evidence was weaker than in the Eight Sets group may be to be an artifact of spuriously low endorsement rates for NONCONF-UNATT words in the Eight Sets group. In conclusion, participants continued applying the generalization to unattested consonant pairs even after ample evidence that only certain consonant pairs can appear in the language.

4 Experiment 2b: Ruling out an identity bias

We interpreted our participants' preference for identical items after one exposure set in Experiment 2a (the One Set group) as reflecting the learning of a probabilistic generalization that held of the exposure words. Before being confident in this interpretation, however, we must rule out the possibility that the preference for identical test items was due to prior bias favoring words with identical consonants rather than 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 biases.

Experiment 2b was designed to rule out a pre-existing preference for words with identical consonants. Participants were exposed to words containing eight consonant pairs, all of which were non-identical. Since the question of interest relates to pre-existing bias, the only relevant exposure group is the One Set one. After the exposure phase, participants provided acceptability judgments on the same unattested items as in Experiment 2a (both CONF-UNATT and NONCONF-UNATT). If participants still showed a preference for identical over non-identical items, despite not having seen any identical items in exposure, this would be evidence that the preference is due to prior bias in favor of identical items. We refer to this hypothesis as the bias hypothesis. If, on the other hand, participants showed no identity preference, the interpretation of the identity preference in Experiment 2a as being due to learning would stand; we refer to this as the learning hypothesis.

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	nuʃe	ʃipu
dʒami	dʒami	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 exposure and testing words. As in Experiment 2a, half of the test words were attested in exposure and half were not. All of the attested words in testing had non-identical consonants. 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 cooccurences in the

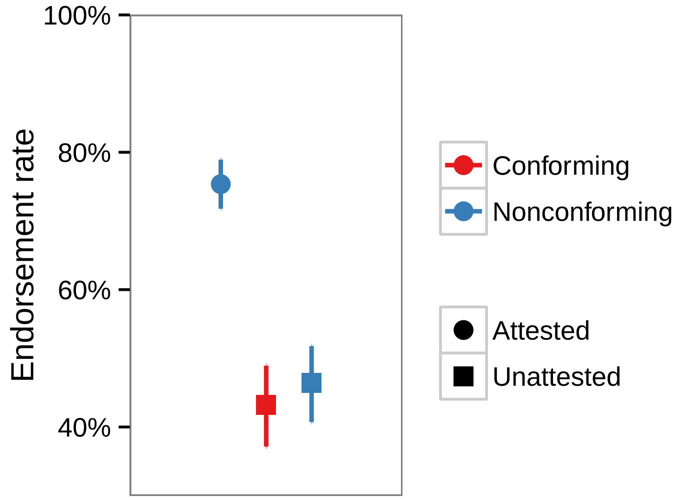

Figure 3: Mean endorsement rates for Experiment 2b. Error bars represent bootstrapped 95% confidence intervals.

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 cooccurence 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.

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 pair and for subject and a random slope by subject for consonant type.

4.2 Results

The results of Experiment 2b are shown in Figure 3. Contrary to the predictions of the bias 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 highly 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 3, the difference between NONCONF-ATT on the one hand and CONF-UNATT and NONCONF-UNATT on the one hand (i.e., the two UNATT conditions collapsed together) was highly significant ($\chi^2(1) = 27.4$, $p < .001$). By contrast, the difference between CONF-UNATT and NONCONF-UNATT did not approach significance ($\chi^2(1) = 0.57$, $p = .45$).

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, the interpretation of the main result of Experiment 2a remains unchanged: Participants showed evidence of learning the broad generalization about identical consonant pairs before learning narrow generalizations about the specific attested consonant pairs.

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 generalization interferes with the learning of narrower generalizations. In Experiment 2a, the presence of the identity generalization prevented learners from attending sufficiently to the narrower generalizations 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 bias to give "yes" responses to 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 bias in favor of 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 bias, and at the same time that the lack of preference for identical items in Experiment 2b was due to learning that offset that bias.

5 Experiment 3: Generalization from a single type

Participants in Experiments 1 and 2a showed evidence of learning an abstract phonotactic generalization before they showed evidence of learning narrower, segment-specific ones. This argues against the specific-to-general approach to generalization: Learners do not need to first learn two or more special cases of a generalization before they can abstract away from these instances and form a generalization. Alternative explanations are still open, however. Participants in the One Set group of Experiments 1 and 2a may have learned only some of the specific types that support the generalization (e.g., two of the four CONF consonant pairs in Experiment 1), enough to form a robust bottom-up generalization, but not enough to produce a statistically significant difference between the endorsement rates for ATT and UNATT items. Alternatively, it may be the case that participants are not confident enough in their knowledge of specific items to distinguish attested from unattested items when that knowledge is based on a small number of tokens, but still use that knowledge to form bottom-up generalizations (Albright & Hayes, 2003).

Experiment 3 rules out these alternative hypotheses by teaching participants a language in which only a single type supports a generalization. If participants still learn the generalization and apply it to types they did not see in exposure, this will provide additional support to the claim that generalizations do not have to be formed in a bottom-up fashion. Specifically, the exposure set contained only one type of voiceless stop onset (e.g., [p]); participants were tested to see if they endorsed the voiceless stops they had not encountered in the exposure phase (for example, [k] and [t], if [p] was the voiceless stop encountered in the exposure phase). Only two words starting with the voiceless stop were presented in exposure. Six filler words starting with onsets that were neither voiceless nor voiced stops were added to make the learning task more challenging and the exposure period longer. As in Experiment 1, participants rated 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 token was presented of each onset. Based on the results of Experiment 1, we expect participants assigned to the Two Types language to acquire the generalization, but not to distinguish attested from unattested onset types. Finally, the Control language did not include any voiceless stops at all: Participants who were assigned this language were only exposed to six filler words. This language served to examine whether participants had a pre-existing bias for or against voiceless stop onsets.

5.1 Method

5.1.1 Materials and procedure

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 FILLER 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 FILLER words, two starting with each of the onsets [w], [y] and [l]. Participants who were taught the Control language were only exposed to these six control 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

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

(a)

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

(b)

Exposure	Test	
FILLER	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.

a different CONF onset ([p] and [t], [p] and [k], or [t] and [k], counterbalanced across participants), in addition to FILLER words (see Table 4b). All participants received a single exposure set.

Exposure words with FILLER onsets were included to make the exposure phase longer; without them the exposure phase of the Single Type language would have been only two words long. Approximants such as [w], [y] and [l] are considered to be voiced, though they are neither stops nor fricatives (Hayes, 2011). If anything, they 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 despite rather than because of the FILLER onsets.

In the test phase, all participants rated 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, didn't provide any evidence for the generalization. Since none of the languages had NONCONF onsets in the exposure phase, NONCONF onsets were always unattested (NONCONF-UNATT). The exposure phase of the Control language didn't 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 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 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.

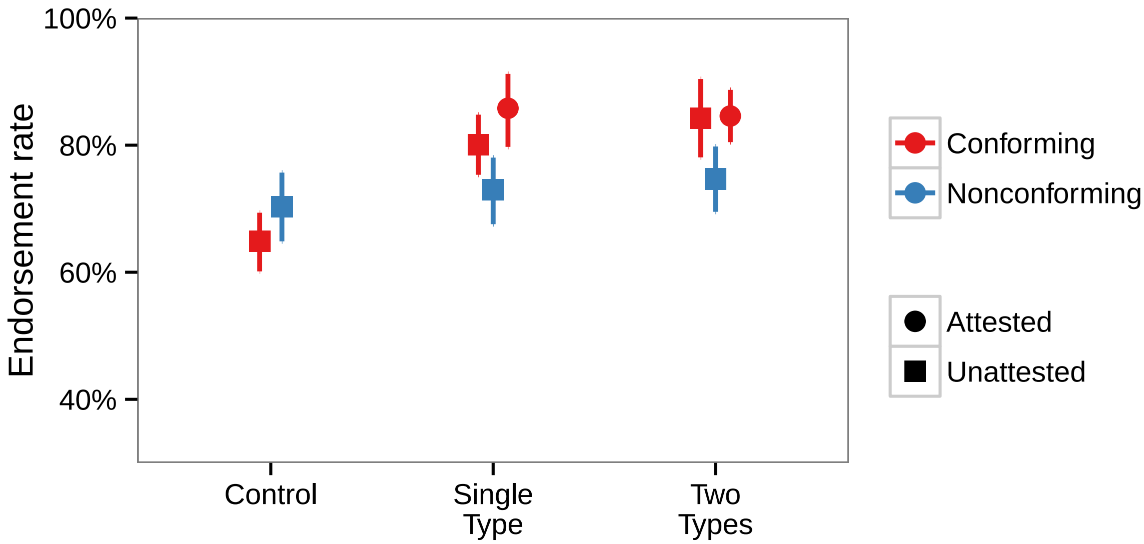

Figure 4: Mean endorsement rates for Experiment 3. Error bars represent bootstrapped 95% confidence intervals.

5.2 Results

Figure 4 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 language, 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 Excluding test words with CONF-ATT onsets

The two main effects were not significant (language: $\chi^2(2) = 1.74$, $p = .42$; onset type: $\chi^2(1) = 1.16$, $p = .28$), but the interaction between language and onset type was significant ($\chi^2(2) = 13.32$, $p = .001$). 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 trend in the opposite direction ($\chi^2(1) = 2.68$, $p = .1$).

The significant simple effect in the Single Type language confirms that learners generalized based on a single CONF onset type in exposure. The reverse trend in the Control language may reflect a tendency to interpret the approximant FILLER onsets in exposure as providing support for voiced over voiceless onsets.

5.2.2 Excluding the Control language

The main effect of onset category was significant ($\chi^2(2) = 12.58$, $p = .002$); the main 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 ($\chi^2(1) = 8.03$, $p = 0.005$) and did not interact with language ($\chi^2(1) = .39$, $p = .53$). There was no significant difference between test words with CONF-ATT and CONF-UNATT onsets ($\chi^2(1) = 1.55$, $p = .21$), and again no interaction with language ($\chi^2(1) = 1.22$, $p = .27$).

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 were indistinguishable ($\chi^2(1) = 0$, $p = .96$); conversely, there was a trend towards higher ratings for CONF-ATT than for CONF-UNATT onsets within the Single Type language ($\chi^2(1) = 2.79$, $p = .09$).

5.3 Discussion

Specific-to-general models predict that generalizations cannot be formed until at least two specific instantiations of the generalization have been encountered. Experiment 3 tested this prediction by exposing participants to the Single Type language, in which two tokens of a single type of voiceless stop onset—e.g., [p]—supported the generalization that onsets can be voiceless stops. Contrary to the prediction of specific-to-general models, participants generalized to unattested generalization-conforming onsets (CONF-UNATT; e.g., [t]), preferring them over NONCONF-UNATT onsets such as [z].

The Control language was designed to rule out two interpretations of the preference that participants who learned the Single Type language showed 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 was due to the presence of six approximant FILLER onsets (though this scenario is unlikely given the absence of any shared phonological features between approximants and voiceless stops). After exposure to this language, which included only FILLER onsets, participants rated test words with CONF-UNATT onsets *lower* than ones NONCONF-UNATT onsets, suggesting that the pattern of endorsement rates for the Single Type language were indeed due to generalization from the single type of voiceless stop in the exposure phase.

In a third language, the Two Types language, the generalization 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 clear evidence of a preference for CONF-ATT over CONF-UNATT in the Single Type language either, though there was a suggestive numerical difference. 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 numerical trend in the Single Type language.

6 General discussion

Speakers generalize their phonotactic knowledge from the sounds attested in their language to sounds from the same phonological classes as the attested ones. If words often end with voiceless stops [p] and [k], speakers will judge words that end with other voiceless stops (e.g., [t]) to be acceptable. The experiments presented in this paper investigated the process by which such generalizations are acquired, focusing on evaluating whether learning follows a specific-to-general order: Do multiple specific instantiations of a generalization need to be acquired before the generalization can be formed?

In Experiments 1 and 2a, participants were divided into groups that received varying amounts of exposure to an artificial language. The consonants of the words in each of the languages supported a phonotactic generalization, but not all possible consonant patterns that conformed to the generalization were presented to participants. In both experiments, participants learned the abstract generalization following minimal exposure: When asked to rate novel words for acceptability, participants in all groups judged words that conformed to the generalization to be better formed than words 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 learning a generalization (e.g., that consonants are often identical) before they showed evidence of learning any of its specific instances (e.g., that [p, p] is a valid consonant pair).

In both Experiments 1 and 2a, the generalization that participants learned was supported by multiple types. In the critical condition of Experiment 3, by contrast, participants were only exposed to a single type of consonant which represented a phonological class. Even when the amount of exposure to the generalization was thereby reduced to the absolute minimum, participants still generalized to other members of that class.

The finding that participants can generalize from a single type is most naturally consistent with simultaneous learning models, in which learners can form generalizations about classes of sounds at the same time as (or before) they acquire knowledge about specific sounds from those classes. Maximum entropy models are a prominent implementation of the simultaneous learning view (Hayes & Wilson, 2008; Pater & Moreton, 2012). In these models, the well-formedness of a sound is derived from a linear combination of weights associated with each of the phonological classes that contain the sound. The well-formedness of a [b] is determined by both the weight for [b] and the weight for the class of voiced stops (as well as 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 will cause both weights to be increased. Even if the only token the learner was exposed to was a [b], then, the learner will rate a novel voiced stop such as [g] as better than a voiceless one such as [k].

After a single exposure set in Experiments 1 and 2a, as well as in the Single Type and Two Types languages of Experiment 3, there was no statistically significant difference between sounds that conformed to the broad generalization and were attested in exposure and those that conformed to the generalization but were not attested in the exposure phase; by contrast, the difference between conforming to nonconforming sounds as a whole was consistently significant. While maximum entropy models predict rapid generalization to unattested conforming sounds, they do not predict that all conforming sounds should be judged as equally well formed. It is true that a single exposure to a [b] would lead a maximum entropy learner to increase some of the weights that apply to [d] (e.g., the weight for voiced stops); at the same time, weights that apply to [b] but not to [d], such as the weight for labials or a weight specific to [b], would also be increased. Consequently, conforming attested sounds (here [b]) would be preferred to conforming unattested ones ([d]). The prediction of both a generalization *and* and an attestation effect is consistent with the empirical endorsement rates after multiple exposure sets, but is inconsistent with the pattern that emerged after minimal exposure (with the usual caveats about the interpretation of negative results).

The absence of an attestation effect after limited exposure can be characterized as a *general-to-specific* learning pattern: generalizations over classes of sounds emerge before generalizations about specific sounds. As Linzen and O'Donnell (2015) show, the general-to-specific empirical pattern does not necessarily reflect an explicit general-to-specific learning strategy. Instead, it can emerge from a simultaneous learner that incorporates a parsimony bias encouraging the learner to represent the input using a single generalization (cf. Chomsky & Halle, 1968, p. 337). For example, after exposure to five different types of voiced onsets (as in Experiment 1), this bias may lead participants 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 learners receive more exposure to the language, however, the absence of conforming unattested sounds becomes more apparent, and prompts learners to revert to a less parsimonious but more accurate sound-specific representation.

The rapid generalization results of Experiment 1 and 2a can be reconciled with the specific-to-general view if we assume that learners always record the identity of the specific sounds they have been exposed to, but avoid deploying that knowledge if the number of exposure words that contained that sound was below a certain threshold. Under this assumption, knowledge about specific sounds might give rise to generalizations about classes of sounds, but would not necessarily lead to a difference in acceptability between attested and unattested sounds. While existing specific-to-general models of phonotactics do not incorporate this assumption—in both the Minimal Generalization Learner (Albright, 2009) and StaGe (Adriaans & Kager, 2010), a generalization cannot have a stronger effect than all of its specific instances—these models can be modified to include a component that discounts knowledge based on fewer tokens (Albright & Hayes, 2002; for simulations, see Linzen & Gallagher, 2014).

It is harder to see how the specific-to-general view could be reconciled with the single-type generalization pattern observed in Experiment 3. Indeed, it seems implausible that a learner that has been exposed to one or two words that start with a [b] will conclude that all words in the language start with a [b] until shown evidence to the contrary. One could imagine a model where upon hearing a word that starts with a particular sound the learner instantaneously constructed all generalizations that contain that sound, of course, but that learner would be empirically indistinguishable from a simultaneous one.

The specific-to-general assumption in phonotactic learning is an example of a broader conservative generalization strategy that is often attributed to language learners. The conservative generalization assumption is motivated by the so-called "subset problem". The form of the argument is as follows (Dell, 1981). The goal of the learner is to characterize the grammar that generated the input it is observing. It can only use positive evidence (i.e., attested forms): No one tells a learner of English that **mpepm* is phonotactically illegal. 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, all words start with a 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 selects Grammar 1, and later on encounters a word that starts with a voiceless stop (e.g., [k]), the learner can revise its decision and assume the wider Grammar 2 instead. The reverse decision is argued to be impossible because of the absence of negative evidence: A learner that chose Grammar 2 would never receive evidence that the generalization was too wide.[5] To avoid overly broad generalizations, 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 subset problem only arises in a specific theoretical model of language acquisition, however (Gold, 1967; Berwick, 1985), and its practical relevance has been called into question (Clark & Lappin, 2011). In particular, while it is true that learners rarely (if ever) receive direct evidence that certain sound combinations are impossible in their language, they often 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 started with [b] and one that started with [d], the learner might conclude that the best characterization of the phonotactics of the language is that all words start with voiced stops. As the learner receives additional exposure to language, however, the systematic absence of [g] onsets becomes more and more conspicuous. If words could start with any voiced stop, the absence of this particular voiced stop would become a "suspicious coincidence" (Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007). This constitutes indirect negative evidence that may cause the learner to revert to a narrower hypothesis, obviating the need for the Subset Principle. This strategy is implemented in the probabilistic model proposed by Linzen and O'Donnell (2015).

Our findings are likely to be applicable outside the domain of phonotactic learning proper. The pronunciation of morphemes often depends on the phonological environment surrounding them. The

[5]In fact, under the assumption that simpler grammars—grammars that can 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).

regular English past tense morpheme *-ed*, for example, has three different variants (allomorphs), [ɪd], [t] and [d]. The choice of allomorph in each case is determined by the final sound of the verb: [ɪd] is used after [t] and [d] (*tended* [tɛndɪd]), [t] after voiceless consonants (*wrapped* [ɹæpt]), and [d] elsewhere (*banned* [bænd]). Speakers use these phonological generalizations to form the past tense of novel verbs: even a speaker who is unfamiliar with the verb *schlep* will agree that its past tense is formed with the [t] rather than the [d] allomorph. This also applies to novel sounds: an English speaker who pronounced the last consonant of *Bach* using the original German [x], which is a voiceless consonant, would likely form the past tense of a putative verb *out-Bach* as [bɑxt] rather than [bɑxd] or [bɑxɪd] (Pinker, 1999).

It is natural to assume that generalizations related to allomorph selection are formed in the same way as the phonotactic generalizations discussed in the current paper (Albright & Hayes, 2003; Gouskova, Newlin-Łukowicz, & Kasyanenko, 2015). The predictions made by specific-to-general models in the case of limited exposure to the language are perhaps even less plausible in the morphological context than in phonotactics. Consider, for example, a learner of English who has only been exposed to the past form of a handful of words that happen to all end with [k], and therefore form their past form with a [t] (e.g., *kicked* and *talked*). This learner would be completely stumped as to the past form of *stop*: it wouldn't even be able to make a guess, since no generalization would have been formed (for a similar point, see Kapatsinski, 2014, p. 16). The results of the present study lead us to predict that learners will show simultaneous rather than specific-to-general learning in the allomorphy scenario as well.

This paper has advocated for a learning procedure that includes phonological classes at all levels of generality as part of the learner's hypothesis space. The existence of phonological classes in the learner's hypothesis space raises the question of the origin of those classes: How do learners know that [b], [d] and [g] form the class of voiced stops, but [b], [s] and [m] do not constitute a phonological class to the exclusion of other sounds? Various answers have been proposed to this question. The inventory of phonological features may be innate (Chomsky & Halle, 1968; Hale & Reiss, 2003); phonological features may emerge as learners group together sounds that sound similar to each other or are articulated in a similar way (Lin & Mielke, 2008); finally, they may simply reflect groupings of sounds according to their behavior in phonological alternations (Mielke, 2008). Our results do not bear on this debate, for two reasons. First, this debate concerns the origin of phonological features in infants. Participants in our experiments were all English speakers, and all of the sounds in the artificial languages they learned were drawn from the English inventory. Consequently, their hypothesis was likely based on the features that are in active use in English phonology.

Second, the question of the origin of phonological features is orthogonal to the distinction between specific-to-general and simultaneous learning more generally. In order to form a generalization over two specific sounds, say [b] and [g], the learner must already be able to represent the fact that the two sounds belong to the same phonological class (specifically, voiced stops); this is exactly the same knowledge that a simultaneous learner would require to analyze a single instance of [b] as a voiced stop (Hale & Reiss, 2003). All extant models of generalization in phonotactics, then, assume that the learner is equipped with the ability to represent phonological classes. It is conceivable, of course, that infants would exhibit two distinct developmental stages, one in which statistics about specific segments were accumulated, and another in which generalizations to wider phonological classes were formed. The empirical evidence to date points in the opposite direction, however (Saffran & Thiessen, 2003; Cristia & Peperkamp, 2012). In reality, the acquisition of phonological features, phonotactics and lexical knowledge may well be intertwined (Feldman, Griffiths, Goldwater, & Morgan, 2013).

All of our participants were English-speaking adults. As such, our experiments are 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 are convergent with results from 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). In another experiment, 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.

7 Conclusion

This paper reported on a series of artificial language experiments of phonotactic generalization in artificial language learning. The experiments showed that participants can learn both segment-specific and abstract phonotactic patterns in an artificial language following a very short exposure session. Abstract patterns (e.g., word onsets are voiced) were learned more quickly than segment-specific ones (e.g., [b] and [d] are valid word onsets); this applied regardless of whether the abstract pattern was categorical or probabilistic, and of whether it was based on a phonological class or on an identity relation across segments. Finally, abstract patterns were acquired on the basis of a single example. We conclude that humans are not conservative generalizers; consequently, models of phonotactic learning should not be based on the assumption that knowledge about specific items is a prerequisite for the formation of abstract generalizations.

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.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., & Bolker, B. (2014). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 0.999999-0)

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.

Berent, I., Shimron, J., & Vaknin, V. (2001). Phonological constraints on reading: Evidence from the Obligatory Contour Principle. *Journal of Memory and Language*, *44*(4), 644–665.

Berent, I., Wilson, C., Marcus, G., & Bemis, D. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry*, *43*(1), 97–119.

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

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

Clark, A., & Lappin, S. (2011). *Linguistic nativism and the poverty of the stimulus*. Malden, MA: Wiley-Blackwell.

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.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

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.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, *77*, 91–106.

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

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.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Gouskova, M., Newlin-Łukowicz, L., & Kasyanenko, S. (2015). Selectional restrictions as phonotactics over sublexicons. *Lingua*. (Retrieved from http://www.nyu.edu/projects/gouskova/downloads/gouskova_newlin-lukowicz_kasyanenko_2015.pdf)

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, *5*, 162–181.

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.

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.

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

Lin, Y., & Mielke, J. (2008). Discovering place and manner features: What can be learned from acoustic and articulatory data. *University of Pennsylvania Working Papers in Linguistics*, *14*(1), 214–254.

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.*

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

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.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford and New York: Oxford University Press.

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.

Pater, J., & Moreton, E. (2012). Structurally biased phonology: Complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad*, *3*(2), 1–41.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

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.

Rose, S., & King, L. (2007). Speech error elicitation and co-occurrence restrictions in two Ethiopian Semitic languages. *Language and Speech*, *50*(4), 451–504.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Devel-*

opmental Psychology, *39*(3), 484–494.

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

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

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, *139*(4), 792–814.

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