There is a simplicity bias when generalizing from ambiguous data*

Karthik Durvasula†

Adam Liter†

Word count: 15,850 (including abstract, figures, tables and references)

Abstract

How exactly do learners generalize in the face of ambiguous data? While there has been a substantial amount of research studying the biases that learners employ, there has been very little work on what sorts of biases learners employ in the face of data that is ambiguous between phonological generalizations with different degrees of simplicity/complexity. In this article, we present the results from 3 artificial language learning experiments that suggest that, at least for phonotactic sequence patterns, learners are able to keep track of multiple generalizations related to the same segmental co-occurrences; however, the generalizations they learn are only the simplest ones that are consistent with the data.

*We would like to thank the Associate Editor and three anonymous reviewers for helping improve this manuscript tremendously. We would also like to thank the audiences at the 2015 Annual Meeting on Phonology, the LSA 2016 Annual Meeting, and the 2016 North American Phonology Conference as well as the Phonology/Phonetics group at Michigan State University for helpful discussions. Moreover, many thanks to Mina Hirzel for recording the stimuli used in our experiments. Additionally, we would like to thanks Russ Werner and Mike Kramizeh at Michigan State University for help with technical matters. And finally, we'd like to acknowledge that Adam Liter was supported by the NSF NRT award DGE-1449815 during portions of the writing and revising of this paper.

†Both authors have contributed equally to this paper and share first authorship.

1 Introduction

Natural language data that children and adults learn from almost invariably have multiple competing generalizations. Even when one narrows their focus to only a particular segmental co-occurrence pattern, it becomes immediately apparent that there are multiple possible generalizations that are consistent with the pattern. Of course, this is not a novel insight, and it is a question that many have grappled with in the past. However, the observation does raise the question that is of primary importance to this article. That is, how exactly do people generalize in the face of data that is ambiguous between generalizations of varying simplicity/complexity? The experiments presented in this article suggest that, at least for phonotactic sequence patterns, learners are able to keep track of multiple generalizations related to the same segmental co-occurrences; however, the generalizations they learn are only the simplest ones that are consistent with the data, where 'simplest' is defined as the use of the fewest representational primitives needed to state the generalization.

With regard to phonotactic patterns, there is considerable evidence that both infants and adults possess phonotactic knowledge of their native languages. This knowledge has been probed through phonotactic judgments of nonce words (Jusczyk et al. 1993; Scholes 1966), through word-segmentation tasks (Friederici and Wessels 1993; McQueen 1998), and through perceptual illusions (Dupoux et al. 1999; Kabak and Idsardi 2007). There is also evidence that this phonotactic knowledge involves not just segmental patterns, but also more abstract, natural class or featural patterns (Albright 2009; Moreton 2002).

The paradigm of artificial language learning (ALL) has been especially fruitful in probing the kinds of patterns that both children and adults learn in the domain of phonotactic learning.[1] There is again substantial evidence now that exposure to words with particular patterns during a training phase of an ALL experiment is sufficient for children and adults to generalize the patterns to novel words (Chambers et al. 2003). Results from such experiments also suggest that, in line with typological asymmetries, people are able to employ 'substantive biases' in generalizing from the

[1] In what follows, we only cite representative work related to phonotactic learning; for a more general review of ALL across a variety of linguistic domains, see Culbertson (2012) and Folia et al. (2010), and, for a more general review of the ALL literature on phonological learning, see Moreton and Pater (2012a,b).

training data (Becker et al. 2011; Moreton 2008; Wilson 2006). Beyond substantive biases that are in line with typological asymmetries, learners in such experiments also appear to exhibit what might be called 'structural biases'; *i.e.*, they exhibit different phonotactic learning biases over different representations (Bergelson and Idsardi 2009; Chambers et al. 2012). Finally, in line with what has been observed for native language phonotactics, learners in ALL experiments also seem to learn featural generalizations; *i.e.*, they are able to access more abstract generalizations than segment sequence generalizations (Cristià et al. 2011; Finley and Badecker 2009). One common theme to much of the research on biases that we just discussed is that they are more directly related to substantive issues, either regarding typological asymmetries, or representational issues.

Given that ALL experiments show such similarity to research on native language phonotactics, they provide a nice alternative, beyond modeling, to understand formal inductive biases employed by a learner during the acquisition of phonotactic patterns. Here, too, there is some recent work probing the issue. Research suggests that, for both adults and children, simpler (particularly, single-feature) generalizations are easier to learn than more complex generalizations (Cristià and Seidl 2008; Kuo 2009; Pycha et al. 2003; Saffran and Thiessen 2003). Others have focused on issues of formal computational complexity and have shown that (adult) learners are able to learn patterns that belong to only a subset of the patterns describable by finite-state automata, labeled the *Subregular* classes (Lai 2015; McMullin 2016).

As mentioned at the outset, our interest in this article is determining what kind of bias, if any, learners impose on data that is consistent with multiple phonotactic generalizations. However, none of the experimental results that were very briefly reviewed above—*i.e.*, neither the experimental results showing that patterns mirror typological tendencies nor those showing that simpler generalizations are learned better than more complex generalizations—directly address the question of what learners learn in the face of ambiguity. Furthermore, while there are quite a few viewpoints espoused for theoretical or logical reasons in the literature, there is far less experimental work on the issue. We describe the few experimental studies that we are aware of that address this question in more detail below, and then we discuss them again in the context of the different theoretical

viewpoints.

Gerken (2006) gave 9 month old infants training stimuli that consisted of syllables that had the pattern AAB (or ABA), *e.g.*, *jidiji* or *jijidi*.[2] Furthermore, in one training condition, the 'B' syllable was always the same syllable *di* (for example, *leledi*, *wiwidi*, *jijidi*, and *dededi*) and, in a second training condition, the 'B' syllable varied between four different syllables *di, je, li, we* (for example, *leledi*, *wiwije*, *jijili*, and *dedewe*). Then they were tested on novel items that were consistent with the AAB (or ABA) pattern (for example, *kokoba* and *popoga*). The infants in their experiments had significantly different looking times, compared to controls, for test stimuli that consisted of AAB (or ABA) syllables with a different set of 'B' syllables, in response to the second training condition but not in response to the first training condition (where all the training words had the same "B" syllable, namely *di*). However, in a follow-up experiment, when the 'B' syllables in the test stimuli were also *di*, then the infants had significantly different looking times even with the first condition. In both experiments, in the first training condition, there were two competing generalizations possible during training; one possible generalization is that all words are AAB (or ABA) and the B is always *di* (the specific generalization), and a second possible generalization is that all the words are AAB (or ABA) (the less specific generalization). Gerken interpreted the results of their experiment as evidence in favor of infants forming the subset (or most specific) generalization when the occasion allowed them to.

A second relevant study is that of Linzen and Gallagher (2014, 2017). The authors were interested in the time-course of generalization, but some of the results in their experiments are relevant to generalizing when the data is consistent with multiple different generalizations. They had a total of four experiments, which are somewhat difficult to explain without an extended discussion; however, we focus on aspects of their results that are crucial for current purposes. For example, in their Experiment 1, they gave one group of participants training words that all started with a voiced obstruent and another group training words that all started with a voiceless obstruent. In the testing phase, they had three types of stimuli: stimuli whose initial consonants were the same as in the train-

[2]We use italics to represent the stimuli here, following Gerken (2006), as we are not sure if these are IPA symbols or English orthography.

ing set (conforming–attested), stimuli whose initial consonants had the same voicing as the training set but were not experienced during the training (conforming–unattested), and stimuli whose initial consonants were inconsistent with the training data (nonconforming–unattested). Overall, across the experiments, Linzen and Gallagher observed that, with sufficient training, participants accepted conforming–attested more than conforming–unattested, which in turn were accepted more than nonconforming–unattested, which means that they are able to learn generalizations but still have a preference for items from the training set. They interpret their results as evidence that, with sufficient exposure, participants learn not only featural (or class level) generalizations but also specific segmental generalizations.

A third relevant study is Cristià et al. (2013), who present some evidence from an ALL experiment that participants do not learn a more complex generalization in the presence of simpler generalizations. As with all ALL experiments, they had a training phase where participants were (auditorily) presented non-words. The test phase stimuli combinations were more complicated than what we present below, but we highlight the crucial aspects of their results. In the test phase, they compared four types of stimuli: old onsets from the training phase ("exposure"), new onsets that had the same phonological features as the narrowest class that described the training set ("within"), new onsets that were one feature away from the narrowest class that described the training set ("near"), and new onsets that were two features away from the narrowest class that described the training set ("far"). Participants rated how frequently they thought the test stimuli had occurred in the training phase. They observed that participants gave the highest frequency ratings to the old onsets ("exposure"); the "within" and "near" received similar frequency ratings, and the "far" received the lowest frequency ratings. Since there was generalization beyond the exposure set of consonants to those outside the set, the results suggest the target grammar was not the subset grammar (*i.e.*, the more complex generalization in our terms). Similarly, the fact that the participants gave similar frequency ratings to "within" and "near" is evidence that the subset generalization was not learned.

Finally, another relevant set of studies are those that suggest local generalizations are privileged

(Finley 2011, 2012; Lai 2015; McMullin 2016). For example, Finley (2011) showed that learners have a bias to learn patterns that are trans-vocalic (also called "first order local"). When presented with training stimuli that contain a version of sibilant harmony across a vowel (*e.g.*, [pisasu], [piʃaʃu]), learners are not willing to extend the pattern to trans-segmental (also called "second-order non-local") occurrences (*e.g.*, [sipasu], [ʃipaʃu]), but learners are willing to extend the pattern from the latter stimuli to the former. This could be argued to show that learners have a bias for the more specific (or complex) generalization, under the assumption that the trans-segmental non-local patterns are simpler to represent than the transvocalic non-local patterns, as the latter patterns make specific reference to ignoring only vowels while the former make no such fine-grained distinction.

Before discussing the different viewpoints about generalization (under ambiguity) and seeing how the above results inform them, we think it is helpful to break down the question of how learners generalize from data that is ambiguous between multiple different generalizations into two subquestions, (1).

1. How do learners learn from ambiguous data?

    (a) Do learners learn just a single generalization that is consistent with a specific segmental co-occurrence pattern, or do they learn multiple (even all) possible generalizations?

    (b) Do learners learn the simplest generalization or the most specific (therefore, most complex) generalization?

The questions might be easier to follow with a concrete example. Assume the exposure stimuli all have consonants that agree in both voicing and continuancy (*e.g.*, [fisu], [pita], [badi], …). There are multiple generalizations that are consistent with the input, including: a Voicing harmony rule (which we will notate as '[α voice]'), a Continuancy harmony rule (which we will notate as '[β cont]'), and a complex Voicing+Continuancy harmony rule (which we will notate as '[α voice, β cont]'). Note, [α voice] here stands for the featural sequence generalization [α voice … α voice], and [β cont] stands for the featural sequence generalization [β cont … β cont];

i.e., both generalizations are over two (albeit, identical) features. Similarly, the conjoined generalization [α voice, β cont] is a featural sequence generalization that involves four features, and therefore counts as a more complex generalization. However, we will use to shorter descriptions throughout to make the text more readable.[3] The first question (1a) asks if the learner learns just a single one of the possible generalizations or if they learn more than one generalization consistent with the data; and the second one (1b) asks if the learner learns the more complex [α voice, β cont] generalization (where the generalization is that both voicing harmony and continuancy harmony have to be present in a stimulus for it to be acceptable), given the data is consistent with the simpler independent generalizations [α voice], [β cont].

There are two important notes worth bearing in mind with regards to the terminology we use. First, throughout, we follow Hayes and Wilson (2008) in employing the notions of *simplicity* and *specificity* to individual rules/constraints instead of whole grammars. Second, we conflate the terms 'most complex' and 'most specific', and use them interchangeably. These are of course separate notions, where the former refers to the intensional description, while the latter refers to the extension set. We maintain this conflation largely because it simplifies the discussion of the previous literature. However, we return to this issue in Section 5 and elaborate on how our results bear on the distinction.

One response to the above questions is to suggest that learners learn just a single, simplest generalization in response to ambiguous data (*i.e.*, just [α voice] or [β cont] in our example above). In one of the earliest discussions on the topic, Chomsky and Halle (1968) and Halle (1961) suggest something similar to this. They suggest that the learner learns the simplest generalization that is consistent with the data, where they define the 'simplest' generalization to be one that uses

[3] First, there are of course other possible generalizations in these stimuli, but we focus on these three. Foreshadowing our experiments, we control for the possibility of other generalizations through the use of Disharmony stimuli. Second, it is not clear if the representations [α, β] count as separate representational primitives by themselves. This is because [α, β] can be thought of as a stand in for the actual feature polarities themselves, which themselves might be unnecessary in a privative feature system. Similarly, the conjunction 'AND' implicit in the description of the complex generalization need not be an explicit element of the intensional description, as its use will depend on the format of the intensional description. In what follows, we consider neither of them for the count of representational primitives, as we do not think that the main argument is affected by their presence in such a count.

the fewest representational primitives (*e.g.*, features, segments, *etc.*).[4] We call this viewpoint the *Simplest Generalization* viewpoint. Throughout this paper, echoing Chomsky and Halle (1968) and Halle (1961), by 'simplicity' we refer to simplicity in terms of representational primitives. Furthermore, we use the word 'simplest' to refer to one extreme on the scale of simplicity, and we use the phrases 'most complex' to refer to the other extreme of the same scale. Therefore, simplicity is based on the *intensional* description and not on the *extensional* sets that result from the description. For example, a generalization that utilizes one feature is simpler than one that utilizes more features; similarly, a generalization that invokes the representation of just a syllable is simpler than one that invokes the representation of a syllable along with that of a segment simultaneously.

A second view that goes in the opposite direction of the *Simplest Generalization* view, in terms of simplicity, is one where the learner keeps track of a single most specific generalization that is consistent with the data (*i.e.*, the complex generalization [α voice, β cont] in our example above). This has been termed the *Subset Principle*[5] (Berwick 1985; Dell 1981). Notably, there is some experimental evidence directly arguing for the *Subset Principle* in the face of ambiguous data. As mentioned above, Gerken (2006) interpreted their results as evidence that infants formed the subset (or most specific) generalization when the occasion allowed them to; however, Cristià et al. (2013) argue against this viewpoint based on their results.

A third view that is very close in spirit to the *Simplest Generalization* idea is the viewpoint that learners learn the simplest generalization, and, in case there is more than one such generalization that can lay claim to being the 'simplest', the learner then keeps track of all such 'simplest' generalizations (*i.e.*, the learner learns both the independent generalizations [α voice] and [β cont], in our example above.). In fact, the Chomsky and Halle (1968) viewpoint discussed above could be charitably extended along these lines, given that they assumed a single 'simplest' generalization in their own discussion. This viewpoint has also been espoused by Hayes and Wilson (2008) in their

[4]Both references explicitly talk about simplicity in the context of phonological features. However, nothing in their discussion, as far as we can see, precludes an extension of the view to other phonological primitives.

[5]However, see Hale and Reiss (2003) for a view that argues that the *Subset Principle* is about lexical representations and not generalizations. Briefly, they suggest that the learner initially posits very specific (thus, richer) lexical representations, and then moves to simpler (or less elaborate) lexical representations with growing experience.

attempt to develop a baseline Maximum Entropy phonotactic learner model. We call this viewpoint the *Multiple Simple(st) Generalizations* viewpoint.

A fourth possible viewpoint is one that suggests that learners are also able to keep track of all the generalizations consistent with the ambiguous data (*i.e.*, the learner learns all the generalizations [α voice], [β cont], and [α voice, β cont], in our example above). However, such viewpoints differ in how they weight the most specific or simplest generalizations. One variation of this viewpoint aligns itself to the *Simplest Generalization* viewpoint, instead of the *Subset Principle*; *i.e.*, learners keep track of multiple (potentially, all) generalizations that are consistent with the data but are biased to weight the simpler generalizations more highly than the more specific ones. We call this view the *Proportional to Simplicity* viewpoint. Although not direct, some evidence for this position comes from the ALL experiments conducted by Linzen and Gallagher (2014, 2017).[6] For their experiments, any increase in acceptability of the conforming–unattested items over the nonconforming–unattested items can be regarded as due to the learning of a simpler pattern involving the feature or natural class of voicing; however, any increase in acceptability of the conforming–attested items over the nonconforming–unattested items could be due to both the learning of a simpler featural generalization and/or a more specific/complex segmental generalization.[7] That is to say, if both the complex and the simple generalizations are learned, then the acceptability of the conforming–attested could be an additive effect of the two generalizations. Overall, across the experiments, they showed that, with sufficient training, participants accepted conforming–attested more than conforming–unattested, which in turn were accepted more than nonconforming–unattested, which means that they are able to learn generalizations but still have a preference for items from the training set. Furthermore, looking carefully at the magnitudes of their results from their Experiments 1 and 2, it seems that, for the largest exposure groups, the difference between the conforming–unattested items compared to the nonconforming–unattested items was

[6]By stating that there is no direct evidence, we do not mean to criticize the results, as this is largely a function of the authors having a different interest/focus than the current paper.

[7]Segmental generalization could be viewed as more complex if segments themselves are not viewed as representational primitives but instead as collections of feature bundles. If, in contrast, segments themselves are seen as representational primitives along with features, then their results are consistent with the *Multiple Simple(st) Generalizations* viewpoint laid out earlier. This is a point we will return to in our interpretation of our own Experiment 1.

larger than the difference between the conforming–attested items and the conforming–unattested items. This thereby suggests that, for these groups of participants, the simpler (featural) generalization had a higher weighting than the more specific (segmental) generalization. The logic behind this potential understanding of their results is further fleshed out below in the context of discussing the predictions of the different viewpoints for our own experiments (see Figure 2).

A second variation of the fourth viewpoint is instantiated in Bayesian models. In a probabilistic formulation of the *Subset Principle*, it has been suggested that learning is proportional to the specificity of the generalization; *i.e.*, a generalization that is more specific is more highly valued or weighted (Linzen and O'Donnell 2015; Tenenbaum and Griffiths 2001; Xu and Tenenbaum 2007).[8] Researchers who support such a viewpoint (typically) take a specific generalization to be a generalization whose extension is a set of possible forms that is closest in size to that of the data encountered through experience (in our case, the training data). We call this viewpoint the *Proportional to Specificity* viewpoint. It needs to be pointed out that some of these claims are made in the context of word-learning, and there is no clear experimental evidence supporting the model's claim for phonotactic sequence learning. Furthermore, while Linzen and O'Donnell (2015) set out to explain the ALL results related to phonotactic patterns of Linzen and Gallagher (2014, 2017) using their model, a crucial prediction of their model—namely, that the weight (or posterior probability) associated with the simplest generalization will decrease with an increasing number of training items—was not observed in Linzen and Gallagher's (2014, 2017) experimental results. So, it is unclear that the experimental evidence from Linzen and Gallagher (2014, 2017) can be interpreted as clear evidence in favor of their model.

These different viewpoints are summarized in Table 1, by way of answering the two different sub-questions laid out above in (1).

In this article, we present 3 ALL experiments that provide evidence for the *Multiple Simple(st)*

[8] We acknowledge the point that Eberhardt and Danks (2011) make that for a Bayesian model to be rational, the model needs to use the generalization with the maximum a posteriori probability; *i.e.*, the model will consistently use the generalization with the highest associated weight. If this is implemented, then such models would make exactly the same predictions as the *Subset Principle*. However, we follow, in our opinion, the intentions of the original papers in assuming that the use of the generalizations is proportional to the (a posteriori) weight assigned to them.

Learn single or multiple generalizations?	Learn simple or most complex generalization?	Viewpoint
Single	Simplest	*Simplest Generalization*
Single	Most specific/complex	*Subset Principle*
Multiple	Simplest	*Multiple Simple(st) Generalizations*
Multiple	Greater weighting for simplest	*Proportional to Simplicity*
Multiple	Greater weighting for most specific/complex	*Proportional to Specificity*

Table 1: Different viewpoints on generalization in the face of ambiguity

Generalizations viewpoint, that learners do learn multiple generalizations in the face of ambiguous data, but the generalizations they learn are only the simplest ones that are consistent with the data; *i.e.*, there is no evidence that learners learn the more complex/specific generalizations in the presence of simpler possibilities. We flesh out more detailed predictions for each of the above viewpoints for each of the experiments after presenting the details of the respective experiments in the following sections. Briefly, in Experiment 1, we look at how participants learn from training words that are ambiguous between two simple featural generalizations ([α voice] or [β cont]), and a more complex featural generalization ([α voice, β cont] satisfied simultaneously). While the experiment presents clear evidence that learners learned the multiple simple generalizations, the evidence for them learning the more complex generalization is confounded, as they could also have simply kept track of a simple generalization over segmental representations. To overcome this confound, in Experiments 2 and 3, we specifically tested participants on stimuli that could not be accepted based simply on the segmental sequences in the training stimuli. The results of Experiments 2 and 3 argue clearly that there is no evidence that the participants kept track of a more complex featural generalization, when simpler generalizations were possible for the training data.

2 Experiment 1

2.1 Methods

2.1.1 Participants

25 English-speaking undergraduates at Michigan State University participated in this experiment for extra credit; however, 2 of the participants were excluded because they always responded 'Yes' to Disharmony and OldStims test items, making it difficult to ascertain whether the participant learned anything at all. Thus, only the data from 23 participants is presented and analyzed in what follows (Mean(Age)=19.9yrs, SD(Age)=1.5yrs; 18 Females & 5 Males).

2.1.2 Materials

In this experiment, participants were trained on a language that consisted of CVCV nonce words. The possible vowels in the language were /a,i,u/, and the possible consonants were /p,b,t,d,f,v,s,z/. All possible CVCV combinations of these vowels and consonants (8*3*8*3=576 items) were recorded prior to running the experiment by a native speaker of American English from Michigan. Having all possible stimuli allowed us to randomize the training and testing stimuli on a participant-by-participant basis at the runtime of the experiment, which was predicated on the hope that by randomizing we would control against any spurious/unintended generalizations in any single stimulus set.

Overall, the experiment consisted of two phases: a training phase and a testing phase. The experiment lasted about 10–15 minutes in total.

Training For the experiments in the paper, we specifically chose to focus on obstruent consonants differing in voicing and continuancy because there are a large enough number of contrasts in English to allow us to have a sufficient number of training and testing stimuli, without there being other concomitant (phonological) featural changes. In the training phase of the experiment, participants were given only CVCV nonce words where the consonants simultaneously agreed in both voicing

and continuancy. For example, [tipa] was a possible word in the language[9] since [t] and [p] are both voiceless and both non-continuants. Similarly, [fisa] was a possible word in the language since [f] and [s] are both voiceless and both continuants. On the other hand, [tisa] was not a possible word in the language since [t] and [s] disagree in continuancy; similarly, [fiza] was not a possible word since [f] and [z] disagree in voicing. The input data was therefore consistent with at least the three following rules: (i) a Voicing harmony rule, (ii) a Continuancy harmony rule, and (iii) a simultaneous Voicing+Continuancy harmony rule. That is, the input data was consistent with multiple generalizations, with different levels of complexity (see Figure 1, below).

The training phase consisted of exposure to 100 possible CVCV nonce words in the language. The set of 100 words that a participant was exposed to during the training phase was chosen randomly from the 144 words in the target language on a participant-by-participant basis using the statistical software R (R Development Core Team 2014).[10]

Each participant was exposed to their list of 100 words twice, in an order that was pseudorandomized by the experimental software, PsychoPy (Peirce et al. 2019), at the runtime of the experiment. The pseudorandomization was constrained in such a way so that a complete pass through the list of 100 words was completed before any repetitions were allowed. The words were presented both orthographically and auditorily on an iMac desktop computer; the auditory presentation occurred through headphones.

Participants were asked to silently mouth the words in order to ensure that they were paying attention to the training items, which would in turn likely facilitate their learning of the language. For a given training trial, participants saw a gray screen with small white writing near the top of the screen that gave the instructions to "Silently mouth the following word"; the orthographic rendition of the word was presented in a larger font in the center of the screen, and the auditory rendition played over headphones. The training trials progressed automatically with an intertrial interval

[9]Henceforth, we will use the term 'language' to refer to the list of all possible words in the training phase; while there were 576 possible CVCV combinations from the segments that we chose for our materials, only 144 of these were such that the consonants agreed in both voicing and continuancy. Thus, when we say 'language' or 'target language' we are referring to these 144 words.

[10]As the the training and testing stimuli lists used for each participant were generated using controlled random seeds, they are fully replicable and the original source files are available upon request.

of 0.5 seconds. The orthographic rendition of the CVCV word was displayed for 1 second; the duration of each trial was therefore equal to either the duration of the auditory presentation of the word or the 1 second duration of the orthographic rendition, whichever was longer. The sound files for the CVCV words were, on average, 0.73 seconds long. For all but 4 sound files, the duration of each trial was 1 second. The remaining 4 sound files had a trial duration of at most 1.03 seconds.

Testing After the training, participants were given a randomized list containing 5 different types of testing stimuli: (i) OldStims, (ii) NewStims, (iii) OnlyVoicing, (iv) OnlyContinuancy, and (v) Disharmony. There were 12 items in each of these 5 categories so that there were 60 test items overall. The participants were asked to determine whether the words they heard were possible words in the language that they had learned in the training phase. The two possible responses were 'Yes' and 'No'. Like with the training stimuli, the testing stimuli were randomly chosen on a participant-by-participant basis.

The OldStims were stimuli that had actually occurred in that participant's training list. The NewStims were stimuli that had *not* occurred in that participant's training list but that did conform to both the voicing harmony rule and the continuancy harmony rule. The OnlyVoicing stimuli had consonants that only agreed in voicing (*i.e.*, they disagreed in continuancy); an example of a possible OnlyVoicing stimulus would be [tisa]. The OnlyContinuancy stimuli had consonants that only agreed in continuancy (*i.e.*, they disagreed in voicing); an example of a possible OnlyContinuancy stimulus would be [zifa]. Lastly, the Disharmony stimuli had consonants that disagreed in both voicing and continuancy; an example of such a test stimulus would be [tiva].

2.2 Predictions

Given that the training data was consistent with both Voicing and Continuancy harmony, there were three different generalizations that were consistent with the data: (a) Voicing harmony ([α voice]), (b) Continuancy harmony ([β cont]), and (c) Voicing+Continuancy harmony ([α voice, β cont]); this is depicted in Figure 1. Note that the third generalization is the most complex and most specific

generalization, while the first two are (equally) simple, where simplicity is defined in terms of representational primitives used for the generalization.

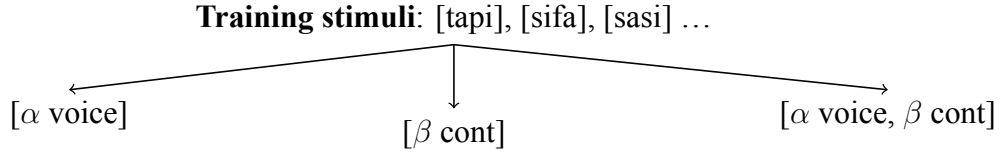

Figure 1: Possible generalizations for training stimuli

Since there are multiple generalizations consistent with the training data, the different viewpoints presented earlier make different predictions about how the learners will generalize from the data, as shown in Figure 2. The figure represents the predicted proportion of 'Yes' responses (averaged over multiple participants) for the different types of test stimuli as per the different viewpoints; the horizontal dotted line represents the baseline proportion of 'Yes' responses for the Disharmony stimuli. There is evidence of learning any of the relevant generalizations only if the proportion of 'Yes' responses to any type of stimuli is above the horizontal dotted line. For example, the OnlyVoicing and OnlyContinuancy stimuli would be above the dotted line only if the learner has learned a Voicing harmony ([α voice]) and a Continuancy harmony ([β cont]) rule, respectively. The predictions related to the different viewpoints are elaborated further below.

The *Subset Principle* would suggest that the learners would only prefer the NewStims and the OldStims compared to the Disharmony stimuli (Figure 2a), as these are the only stimuli that are consistent with the more complex/specific generalization (namely, [α voice, β cont]). The learners should not find the OnlyVoicing and OnlyContinuancy stimuli as more acceptable than Disharmony, as neither of them are consistent with the most specific generalization.

The *Simplest Generalization* viewpoint predicts that some learners will generalize to Voicing harmony ([α voice]), while others will generalize to Continuancy harmony ([β cont]), as depicted in Figure 2b. Therefore, some learners should prefer OnlyVoicing stimuli and others OnlyContinuancy stimuli compared to the Disharmony stimuli. Furthermore, since NewStims are consistent with either generalization, they should be as acceptable *for any single speaker* as either the OnlyVoicing stimuli or the OnlyContinuancy stimuli. As a consequence, when averaged over mul-

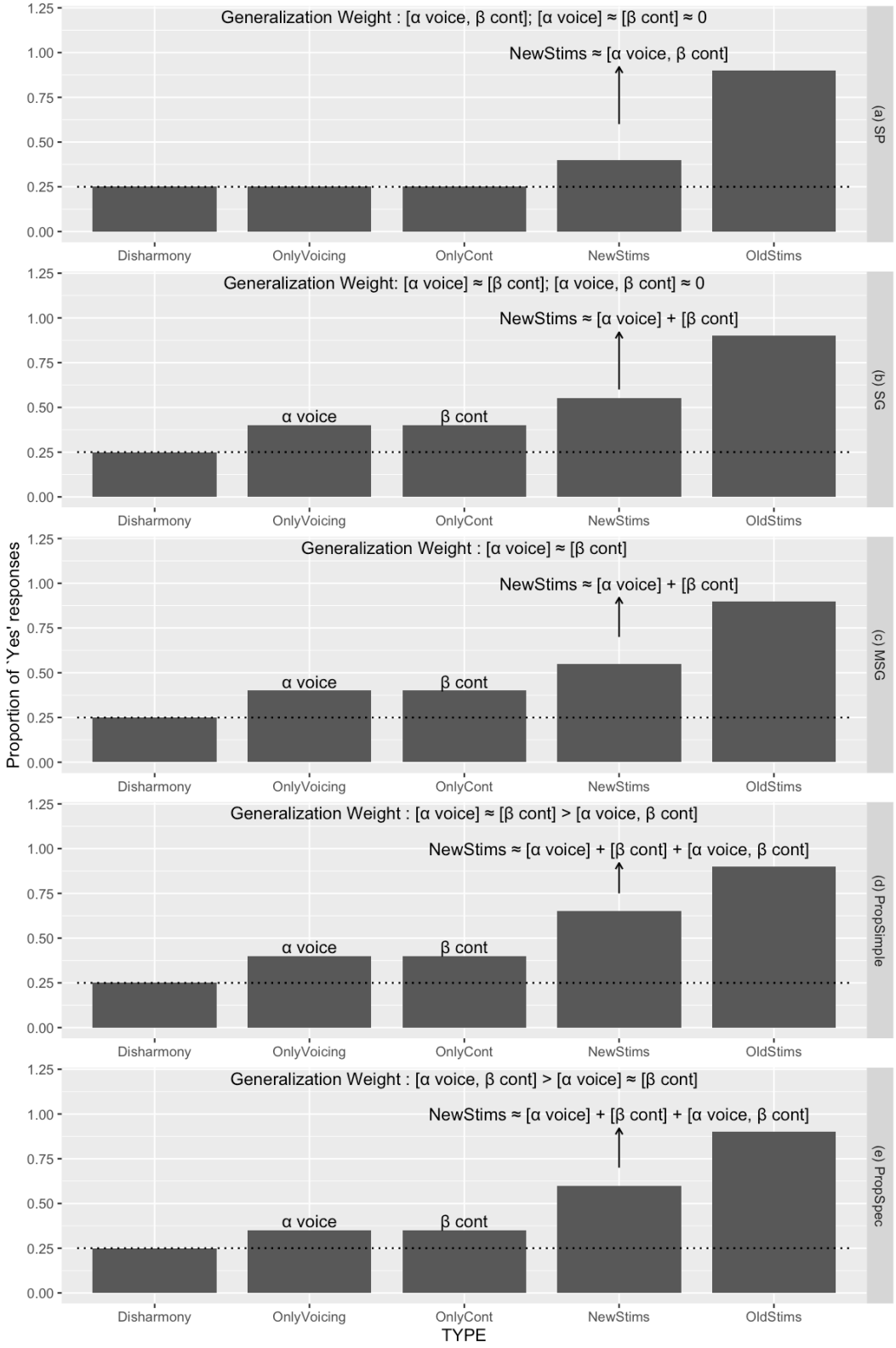

Figure 2: Predicted proportion of 'Yes' responses (averaged over multiple participants) for the test stimuli under different viewpoints (SP = *Subset Principle*; SG = *Simplest Generalization*; MSG = *Multiple Simple(st) Generalizations*; PropSpec = *Proportional to Specificity*; PropSimple = *Proportional to Simplicity*; Dashed line = baseline proportion of 'Yes' responses to Disharmony stimuli).

tiple speakers, the acceptance of NewStims will look as if the acceptance arises from an additive effect of learning both Voicing and Continuancy harmony separately. However, crucially, there is a prediction that there should be a negative correlation between learning Voicing and Continuancy harmony; *i.e.*, as the acceptance for OnlyVoicing stimuli increases, the acceptance for OnlyContinuancy stimuli should decrease, since any given participant should have only learned one of the simple generalizations, not both.

The *Multiple Simple(st) Generalizations* viewpoint predicts that learners learn both Voicing harmony ([α voice]) and Continuancy harmony ([β cont]) as separate generalizations (Figure 2c). Therefore, both the OnlyVoicing stimuli and the OnlyContinuancy stimuli should be preferred over the Disharmony stimuli. As a consequence, there is a predicted additive effect on the NewStims, which are consistent with both Voicing harmony and Continuancy harmony (*i.e.*, NewStims $\approx$ OnlyVoicing + (OnlyContinuancy – Disharmony)). Furthermore, based on the assumption that greater learning of each generalization is driven by greater overall learning (due to performance factors such as more attention to the task, more aptitude for the task, …), there should be a positive correlation between learning Voicing and Continuancy harmony; *i.e.*, as the acceptance for OnlyVoicing over Disharmony stimuli increases, the acceptance for OnlyContinuancy over Disharmony stimuli should also increase.

The *Proportional to Simplicity* viewpoint predicts that learners learn all three generalizations (namely, [α voice], [β cont], and [α voice, β cont]), but the importance given to each of the generalizations is expected to be directly proportional to their simplicity (Figure 2d); therefore, the simpler generalizations (*i.e.*, [α voice] and [β cont]) should be learned better than the more specific generalization ([α voice, β cont]). As a consequence, learners will prefer the OnlyVoicing stimuli and the OnlyContinuancy stimuli over the Disharmony stimuli. Furthermore, since the NewStims are in the extension of all three generalizations, the preference for NewStims should be more than just an additive effect of the preference for the OnlyVoicing stimuli and the OnlyContinuancy stimuli over the Disharmony stimuli; *i.e.*, a super-additive (or interactive) effect is predicted for the NewStims compared to the OnlyVoicing stimuli and the OnlyContinuancy stimuli. Finally, since the weight

associated with the more specific generalization is not as large as those with the simpler generalizations, the contribution of the more specific generalization to the acceptability of the NewStims will also be consequently smaller than the contributions of the simpler generalizations (*i.e.*, [α voice], or [β cont]). Therefore, an interactive or super-additive effect observed for the NewStims over the OnlyVoicing stimuli and the OnlyContinuancy stimuli should be smaller than the preference for the OnlyVoicing stimuli and the OnlyContinuancy stimuli, when compared to Disharmony (*i.e.*, NewStims $\approx$ (OnlyVoicing + (OnlyContinuancy – Disharmony) + X), where X $<$ (OnlyVoicing – Disharmony) and X $<$ (OnlyContinuancy – Disharmony)).

Finally, the *Proportional to Specificity* viewpoint makes similar predictions to the *Proportional to Simplicity* viewpoint, as it too predicts that all three generalizations will be learned (Figure 2e). However, the one difference is in the importance of the weight given to each of the generalizations. Where the *Proportional to Simplicity* viewpoint is biased towards simpler generalizations, the *Proportional to Specificity* viewpoint is biased towards more complex/specific generalizations. So, like the *Proportional to Simplicity* viewpoint, the *Proportional to Specificity* viewpoint predicts that the preference for NewStims should be more than just an additive effect of the preference for the OnlyVoicing stimuli and the OnlyContinuancy stimuli over the Disharmony stimuli (*i.e.*, again a super-additive, or interactive, effect is predicted). However, since the most complex generalization is weighted more than the simpler generalizations, the interactive effect observed should be larger than the preference for either the OnlyVoicing stimuli or the OnlyContinuancy stimuli, when compared to Disharmony (*i.e.*, NewStims $\approx$ (OnlyVoicing + (OnlyContinuancy – Disharmony) + X), where X $>$ (OnlyVoicing – Disharmony) and X $>$ (OnlyContinuancy – Disharmony)).

2.3 Results

A visual inspection of the mean proportion of 'Yes' responses suggests that all four types of test stimuli (OnlyVoicing, OnlyContinuancy, NewStims, and OldStims) are more acceptable to participants than the Disharmony stimuli (Figure 3). Furthermore, the proportion of 'Yes' responses for the NewStims appears to be more than just an additive effect of OnlyVoicing and OnlyContinuancy

stimuli.

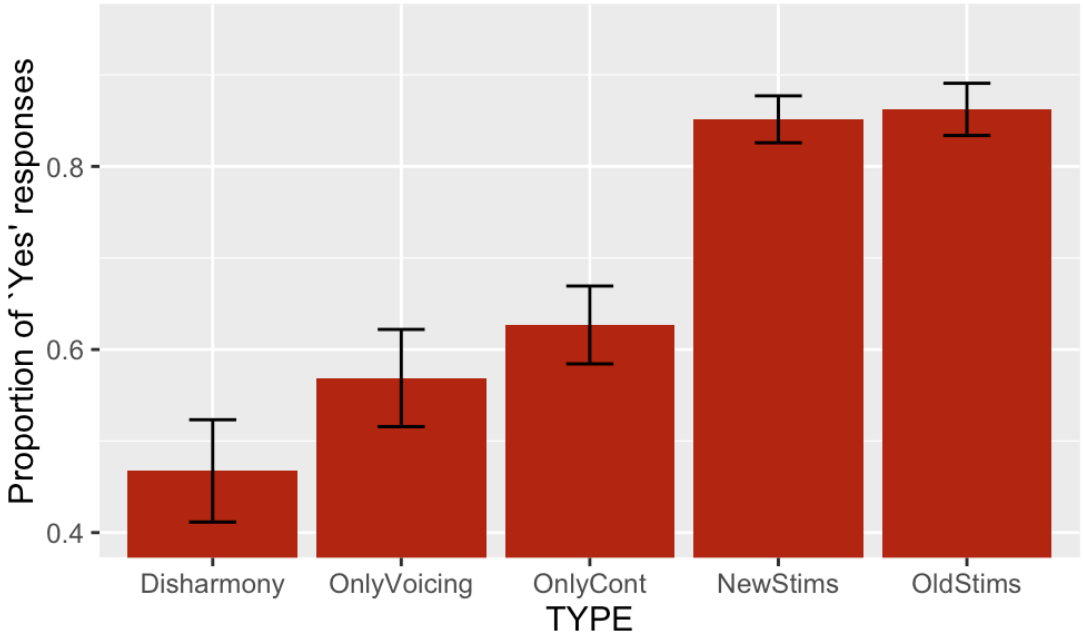

Figure 3: Proportion of 'Yes' responses to test stimuli in Experiment 1 (error bars represent standard errors; OnlyCont = OnlyContinuancy).

In order to confirm the observations made by visual inspection of the results, we conducted a statistical analysis. In this article, wherever possible, participant responses were analyzed using mixed-effects logistic regression models in the statistical software package R (R Development Core Team 2014). The models were fitted using the `glmer` function available through the `lme4` package (Bates et al. 2015). We attempted to obtain the maximal random effects structure that was possible (Barr et al. 2013). However, as is typical in psycholinguistic data (and in our own experience), the models with the most complex random effects structures did not converge. It is important to note that the field and the statistical literature in general have not come to a consensus on how to best proceed in identifying the best random effects structure, especially when a model with a particular random effects structure does not converge (Bolker 2014). In what follows, we describe the random effects structure selection process that we used for our experiments by following other experienced linear mixed-effects modelers in psycholinguistics (Barr et al. 2013; Jaeger 2009, 2011).

We identified the appropriate random effects structure by keeping the fixed effects constant;

we used the full fixed effects model for the experiment (*i.e.*, with interactions for all the fixed effects, if relevant). We started with the most complex random effects structure. In the case of non-convergence of the complex random effects model, we systematically pared down the random effects structure till convergence was reached. The least complex random effects structure we entertained was one with a varying intercept for both subjects and items. When convergence was reached, the corresponding random effects model was identified to be the maximal random effects structure possible for the data. We then performed model comparison (using the maximal random effects structure possible for the data that we identified in the manner just detailed) in order to identify the best combination of fixed effects; specifically, we compared models through backwards elimination of non-significant terms, beginning with the interactions, through a Chi-squared test of the log likelihood ratios. The most complex fixed effects model entertained was the full model with all interaction terms, and the least complex model entertained was the model with only an intercept term and no fixed effects.

Using the above procedure, we attempted to fit logistic mixed-effects models for all the responses in Experiment 1, where the dependent variable was a binary variable that codes for whether participants responded with a 'Yes' or not. To find out if the responses to the NewStims were more than an additive effect of the OnlyVoicing and OnlyContinuancy responses (*i.e.*, a super-additive effect), we coded the OnlyVoicing stimuli as *Voicing*, the OnlyContinuancy stimuli as *Continuancy*, the NewStims as both *Voicing* and *Continuancy*, and the Disharmony stimuli as neither *Voicing* nor *Continuancy*. The random effects structure included a varying intercept for subjects and items. The best model was one with an interaction effect (Table 2). This suggests that the responses to the NewStims cannot be modeled as simply an additive effect of the responses to the OnlyVoicing and OnlyContinuancy stimuli; crucially, the interaction effect is larger than either of the main effects.

We include the comparison between the above model and a model without an interaction term. As can be seen in Table 3, on the basis of the Chi-squared test and the AIC/BIC,[11] the model with the interaction term is the better model.

[11] Note, lower AIC/BIC values are better.

Fixed Effect	Estimate	z-value	Pr($>z$)	
(Intercept)	–0.1231	–0.544	0.2934	
Voicing	0.4758	2.513	0.0059	**
Continuancy	0.7574	3.920	<0.0001	***
Voicing:Continuancy	0.8881	3.032	0.0012	**

Table 2: Best fitting logistic mixed-effects model for Experiment 1

Model	AIC	BIC	ChiSq	Pr($>z$)
Without interaction term	1293	1318		
With interaction term	1285.7	1315.7	9.3	0.002

Table 3: Model comparison with a model without an interaction term

2.4 Discussion

The results of Experiment 1 suggest that participants are able to learn the simpler generalizations even when a more complex generalization is consistent with the training data. This must be true as the OnlyVoicing and the OnlyContinuancy stimuli were both rated higher than the Disharmony stimuli during training, which would not be predicted if participants had only learned the more complex generalization (*i.e.*, [α voice, β cont]). Therefore, the viewpoint that suggests that the learner would learn only the most complex/specific generalization (*i.e.*, the *Subset Principle*) is inconsistent with the results.

Furthermore, the results also suggest that the 'Yes' responses to the NewStims cannot be modeled simply as an additive effect of the two simpler generalizations (*i.e.*, [α voice] and [β cont]). Therefore, the results can be reasonably interpreted as support for the viewpoints that claim that learners keep track of all the possible generalizations; *i.e.*, it can be seen as support for the *Proportional to Simplicity* and *Proportional to Specificity* viewpoints.

However, it is possible that learners keep track of segment (consonant) sequence generalizations along with featural generalizations; *i.e.*, segments, like features, are representational primitives. Note, such a view is independently needed to account for phonological patterns such as segmental metathesis, segment epenthesis, and segment deletion (also see Kazanina et al. (2017) and Albright (2009) for a similar claim that segment-sized representational primitives are needed.)

Furthermore, if consonant sequence generalizations are considered to be as simple as featural generalizations by learners (provided they involve the same number of primitives), then the *Multiple Simple(st) Generalizations* viewpoint would say that learners should be able to keep track of consonantal sequences, separately from their featural content. In such a case, saying [p…t] would be simpler than a conjoined featural generalization [α voice, β cont], because "simplicity" in this paper refers to the *intensional* description, and it is established by counting the number of representations (features, segments, *etc.*) in the generalization. As a consequence, since the responses to the NewStims would then be an additive effect of the Voicing harmony, Continuancy harmony, and the consonant sequence generalizations learned during training, even the *Multiple Simple(st) Generalizations* view would say that the responses to the NewStims would be more than an additive effect of just the responses to the OnlyVoicing and OnlyContinuancy stimuli. Given this potential confound from the possibility that segments themselves can independently be representational primitives, the evidence for the learning of the more complex featural generalization is not clear from Experiment 1.

Given that there is no clear evidence that the learners learned the complex (conjoined) constraint, we cannot adjudicate between the *Proportional to Specificity* and *Proportional to Simplicity* viewpoints directly. Adjudicating between these two viewpoints would be more appropriate to do in the case of Experiments 2 and 3, where the segmental generalization confound isn't present. However, foreshadowing the results, in these experiments, there is *no* evidence that participants learned a more complex featural generalization (*i.e.*, [α voice, β cont]). Thus, adjudicating between the *Proportional to Specificity* and *Proportional to Simplicity* becomes unnecessary, as both viewpoints predict that participants should learn the more complex generalization, contrary to what we find in our subsequent results.

The main findings in Experiment 1 are that speakers are able to keep track of the simple featural generalizations ([α voice] and [β cont]) as evidenced by the fact that they accept OnlyVoicing and OnlyContinuancy stimuli during the test phase. What there is no clear evidence for is if they are learning the more complex (conjoined) featural generalization ([α voice, β cont]), because the

interaction result observed with NewStims could also be explained if speakers use segments themselves as representational primitives to form generalizations. In Experiments 2 and 3, we focus specifically on avoiding the above confound, to see if there is any evidence of participants learning complex (conjoined) *featural* generalization, and we show that the interactive effect found in Experiment 1 for the NewStims disappears when the possibility of using a segmental generalization is removed. This suggests that learners are not learning the more complex (conjoined) featural generalization, when the simpler featural generalizations are present.

3 Experiment 2

In Experiment 1, learners' responses to the NewStims appeared to be more than an additive effect of their responses to the OnlyVoicing and OnlyContinuancy stimuli. However, the NewStims contained consonant sequences that may have been heard during training, so the super-additivity could simply be a result of the learners keeping track of consonant sequence generalizations alongside simple featural generalizations. To control for this possibility, in Experiment 2, we withheld certain pairs of consonants during training and created NewStims during testing using those withheld consonant sequences (as described below in Section 3.1.2). As a consequence, the responses to NewStims can no longer be influenced by any consonant sequence generalizations. Therefore, if the responses to the NewStims are still super-additive over the responses to OnlyVoicing and OnlyContinuancy, that would constitute evidence that, in the face of ambiguous data, the learners are able to keep track of not only the simpler featural generalizations ([α voice] and [β cont]), but also the more complex featural generalization ([α voice, β cont]).

3.1 Methods

3.1.1 Participants

78 English-speaking undergraduates at Michigan State University participated in this experiment for extra credit. We ran many more participants in Experiment 2 to ensure that any lack of a

super-additive effect observed was not due to a lack of power, as discussed further in Section 3.3. The decision to run at least 50 participants was made before, and was not based on data peeking. This was to make sure that there was sufficient statistical power. However, since the experimental participants were recruited through extra credit, we ended up having 78 participants.

Out of the original 78 participants, 15 of the participants were excluded due to non-learning (*i.e.*, they always responded 'Yes' to the Disharmony and OldStims test items). Thus, only the data of 63 participants is presented and analyzed in what follows (Mean(Age)=20yrs, SD(Age)=3yrs; 46 Females & 17 Males).

3.1.2 Materials

The design of Experiment 2 was nearly identical to that of Experiment 1. The possible vowels and consonants were the same, and the experiment also took about 10–15 minutes overall. The only minor differences between Experiment 1 and Experiment 2 are discussed immediately below.

Training The only difference between Experiment 1 and Experiment 2 in the training phase was that we withheld certain consonant sequences in the training phase of Experiment 2. As mentioned before, this was done to address a possible confound in Experiment 1 where participants may have been keeping track of consonantal sequences. The consonant pairs we withheld were randomized on a participant-by-participant basis. For example, one participant would have never received [tVpV] or [pVtV] in their training input while another participant would have never received [fVsV] or [sVfV]. These participants would have never heard these consonant pairs, but they would have nonetheless still heard these consonants in other contexts in their training. For example, while the first participant would never have heard words of the form [tVpV], this participant would have heard words of the form [tVtV] and [pVpV]. We allowed for the possibility of the learners hearing identical consonant sequences consisting of each of the consonants in the withheld consonant sequences to make sure that learners did not choose a 'No' response to the withheld consonant sequences in testing purely because the consonants themselves were novel to them.

Other than this constraint on the possible training stimuli, the training procedure for Experiment 2, including the presentation of the training stimuli, was exactly the same as in Experiment 1.

Testing Like with the training, the testing phase in Experiment 2 was nearly identical to the testing phase in Experiment 1. There were the same 5 different types of testing stimuli, consisting of 12 items each for a total of 60 test items. The only difference was that the NewStims during the test phase of Experiment 2 consisted of words that used the consonant pairs withheld during training.

In other words, the NewStims in Experiment 1 were novel words, but they contained consonant sequences that a participant might have previously heard. For example, if a NewStims test stimulus in Experiment 1 was [fasi], the participant never would have heard [fasi] in training, but the participant might have heard [fisu] in training. This is no longer the case in Experiment 2. In Experiment 2, the NewStims test stimuli were novel words that had novel consonant pairings (because the consonant pairings were withheld during training, as discussed immediately above).

3.2 Predictions

Since the current experiment controls for the possibility of participants using segmental generalization for the test stimuli, the predictions of all the viewpoints presented earlier are the same as discussed in Section 2.2 (see Figure 2).

3.3 Results

A visual inspection of the mean proportion of 'Yes' responses suggests, as in Experiment 1, that all four types of test stimuli (OnlyVoicing, OnlyContinuancy, NewStims, and OldStims) are more acceptable to participants than the Disharmony (Figure 4). Moreover, two further observations can be made. First, the proportion of 'Yes' responses for NewStims seems much lower than in Experiment 1. Second, unlike in Experiment 1, the 'Yes' responses for the NewStims appear to be no more than an additive effect of the responses to the OnlyVoicing and OnlyContinuancy stimuli, over and above the Disharmony stimuli.

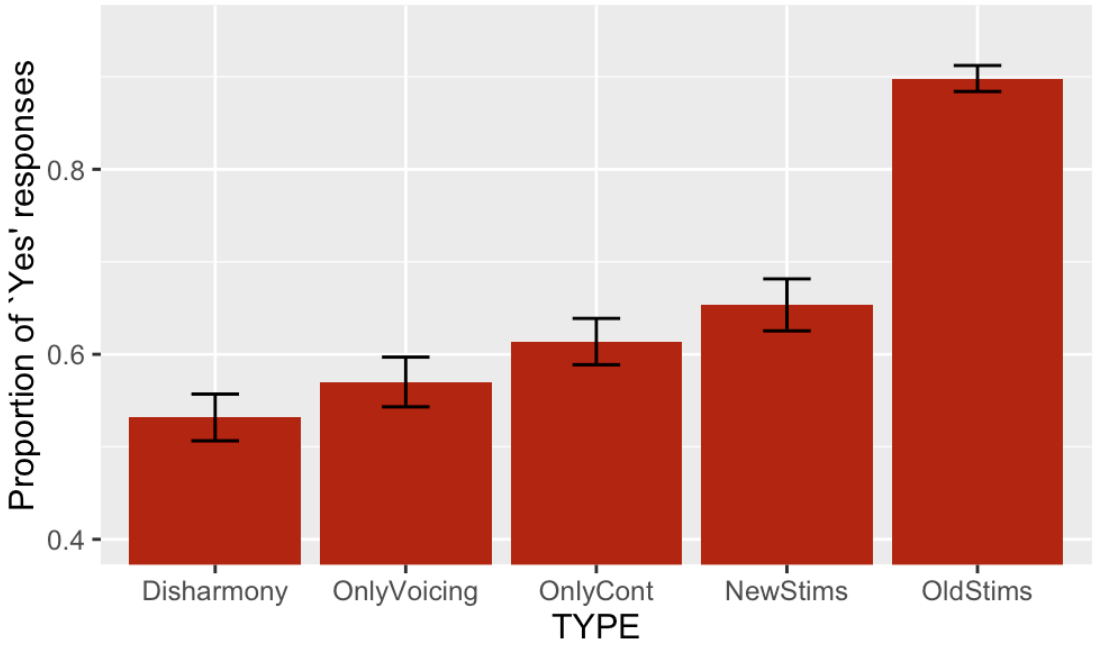

Figure 4: Proportion of 'Yes' responses to test stimuli in Experiment 2 (error bars represent standard errors; OnlyCont = OnlyContinuancy).

As in Experiment 1, to find out if the responses to the NewStims were more than an additive effect of the OnlyVoicing and OnlyContinuancy responses (*i.e.*, a super-additive effect), we coded the OnlyVoicing stimuli as *Voicing*, the OnlyContinuancy stimuli as *Continuancy*, the NewStims as both *Voicing* and *Continuancy*, and the Disharmony stimuli as neither *Voicing* nor *Continuancy*. And again, we attempted to fit a mixed-effects logistic regression model (following the procedure discussed in Section 2.3). The random effects structure, as in Experiment 1, included a varying intercept for subjects and items. The best model was one with two simple main effects (Table 4). The results of the modeling suggest that the responses to the NewStims can in fact be modeled as simply an additive effect of the responses to the OnlyVoicing and OnlyContinuancy stimuli. This is consistent with what can be observed visually in Figure 4. Given that the best model was one with two simple main effects, there appears to be no evidence of a super-additive effect for the NewStims.[12]

In order to better understand if the lack of an interaction effect (Figure 4, Table 4) is either due

[12]The model with the interaction term for *Voicing* and *Continuancy* was not significantly better than the best model. Furthermore, the interaction term was not significant in that model ($\hat{\beta} = 0.015, p = 0.93$).

Fixed Effect	Estimate	z-value	Pr($>z$)	
(Intercept)	0.1625	1.279	=0.2	
Voicing	0.1829	1.941	=0.05	*
Continuancy	0.4031	4.272	<0.001	***

Table 4: Best fitting logistic mixed-effects model for Experiment 2

to the atypical responses of just a few participants or perhaps due to a lack of sufficient statistical power in our experiment, we further plotted the cumulative proportions of 'Yes' responses to each *Type* of test stimulus with increasing number of participants (Figure 5). The plot suggests that the relative differences in the cumulative effect sizes stabilize after about 25–30 participants, thereby suggesting the non-interactivity is not due to a lack of power in Experiment 2. The plot also includes the putative additive effect of OnlyVoicing and OnlyContinuancy (VoicePlusCont = OnlyVoicing + (OnlyContinuancy – Disharmony)). As can be observed, the VoicePlusCont line almost perfectly coincides with the NewStims responses after about 25–30 participants, thereby providing further evidence that the responses to NewStims are indeed no more than an additive effect of the responses to the OnlyVoicing and OnlyContinuancy stimuli.

Finally, we also wanted to take a closer look at the data to see if participants were indeed learning both simple generalizations (*i.e.*, [α voice] and [β cont]). It is possible to read the data presented so far for Experiments 1 & 2 as consistent with some participants learning Voicing harmony, and some others learning Continuancy harmony. That is, since we presented only the overall mean proportion of responses, it is not clear if each participant was learning both simple generalizations. Therefore, to confirm that the participants were really learning both simple generalizations, we looked at the proportion of 'Yes' responses to OnlyContinuancy stimuli and the proportion of 'Yes' responses to OnlyVoicing. Note, we could not make a similar comparison in Experiment 1, as 23 data points (one corresponding to each participant) are usually seen as insufficient to fit a simple linear regression model (Field 2013).[13] If each participant was really learning just one simple generalization (at the cost of the other), then there should be a trade-off in their responses to the

[13]Though we did not include the relevant plot, the reader might be interested in knowing that there was also a significant positive correlation for the same comparison in Experiment 1 ($\hat{\beta} = 0.361, p = 0.03$).

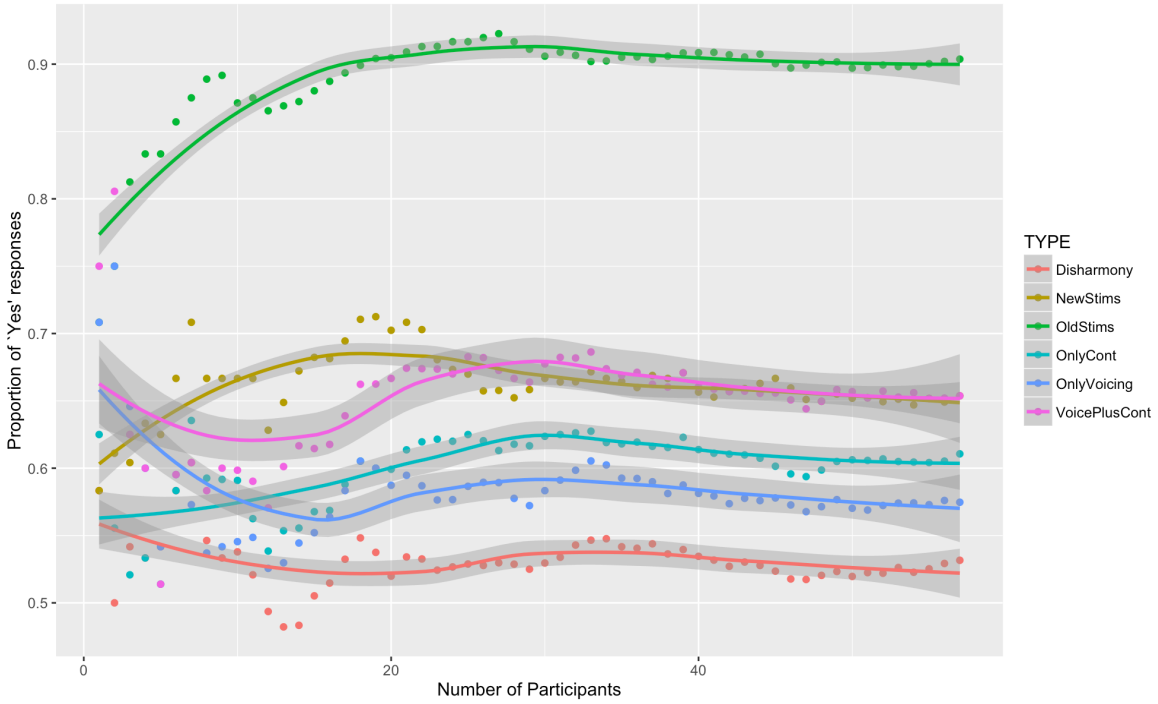

Figure 5: Cumulative proportions of 'Yes' responses to test stimuli with increasing number of participants in Experiment 2 (ribbons represent 95% confidence intervals; OnlyCont = OnlyContinuancy).

OnlyVoicing and OnlyContinuancy stimuli (*i.e.*, there should be a negative correlation between the two responses). On the other hand, if each learner is learning both generalizations, there should be a positive correlation between the responses to the OnlyVoicing and OnlyContinuancy stimuli. In Figure 6, we indeed see a positive correlation ($\hat{\beta} = 0.469, p < 0.00001$). This suggests that if a learner thought that the OnlyVoicing stimuli were more like the training data, they were also likely to think that the OnlyContinuancy stimuli were more like the training data.

3.3.1 Is there any evidence that the complex generalization was learned?

The logistic regression models that we've presented cannot directly test if a model without an interaction effect is supported by the data.[14] The issue is the following: it is possible for the learners to learn a more complex generalization, along with the simpler generalizations, and for

[14]Thanks to the Associate Editor for highlighting this issue to us and for providing a way forward by suggesting the Monte Carlo simulations we present here.

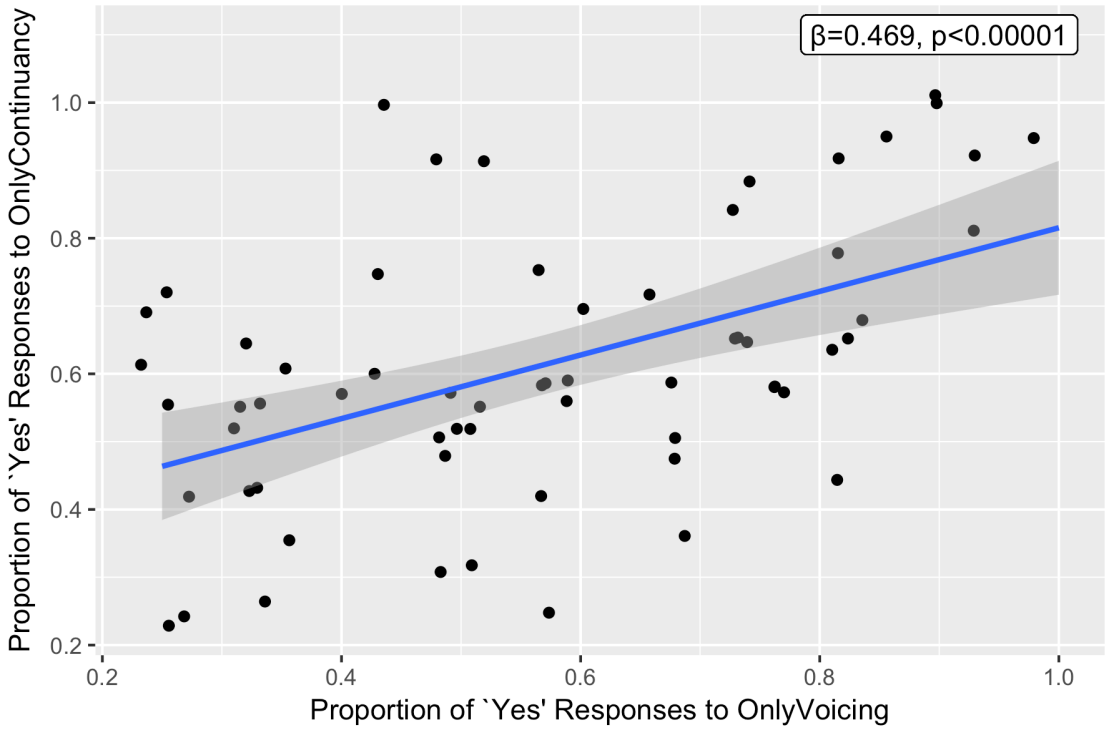

Figure 6: Correlation between increase in 'Yes' responses in Experiment 2 to OnlyVoicing compared to those of OnlyContinuancy (note: jitter has been added to the plot to reveal overlapping points).

the interaction term in the logistic regression to still be non-significant. The fundamental problem is the indirect relationship between interaction terms in logistic regression and super-additive raw probability.

Since the original logistic regression analyses don't quite address the issue we are interested in probing, we did a Monte Carlo simulation. In order to do Monte Carlo simulations, one has to be explicit about the underlying probability models, and about how learners would employ the multiple generalizations that they have learned. There are at least two ways in which one could flesh out the underlying probability models (see Supplementary Materials A for an annotated R script for explicit statements of the models and related simulations):

(a) *Model A:* the learner uses all the generalizations *simultaneously* while making an acceptability judgment. This can be cashed out as a probability that is the product of the probabilities associated with each of the generalizations.

(b) *Model B:* despite knowing multiple generalizations, the learner uses *only one generalization at any one time* with equal probability of picking any of the generalizations, while making an acceptability judgment; *i.e.*, the generalizations are used *individually* and *mutually exclusively* while making an acceptability judgement. This can be cashed out as the average of the probabilities associated with each of the generalizations.

In our opinion, both of these are reasonable probability models, and it is not possible to decide a priori which is the more appropriate model. For this reason, we present the results of the Monte Carlo simulations with each underlying probability model.

The steps we used for the Monte Carlo simulation are as follows:

Step 1: Sketch out a model of the predicted probability of acceptance, with and without the complex generalization. (We ran simulations for both of the probability models fleshed out above, (a) and (b)).

Step 2: Fit underlying probabilities to the observed data, without allowing for any complex generalization. This is the *null hypothesis*.

Step 3: Use the Monte Carlo approach to generate the experimental data for each experiment 1,000 times[15] (replicating the combinatorics of the experiments) from these underlying probabilities and fit a logistic regression model (with interaction) to each iteration. This gives the distribution of interaction coefficients expected under the null hypothesis of no complex generalization.

Step 4: Fit a logistic regression model (with interaction) to the observed data. This gives the actual/observed interaction coefficient.

Step 5: Is the interaction coefficient for the actual data further away from the mean of the interaction coefficients obtained from the simulation than 95% of the interaction coefficients obtained from the simulation?

In Table 5, we present the results based on two sets of 1,000 simulations of the null models for Experiment 2 (simulations were repeated to ensure reliability). We present the proportion of interaction coefficients for data simulated from the null model that were further away from the mean of the interaction coefficients than the actual interaction coefficient observed in the experiments. This is essentially a p-value with the simulations giving us the sampling distribution of the interaction coefficients, and a higher proportion means the interaction is closer to the mean of the interaction coefficients under the null model. Effectively, we have run a two-tailed test. We believe this is more appropriate, as the interaction coefficient could have been both more than or less than the mean of the interaction coefficients for the simulated data from the null model.

As can be seen from the results, there is simply no evidence of a more complex generalization being learned. The model with only the simple generalizations captures the data almost perfectly; this is so, because the interaction effect for the actual data is *very close* to the mean of the interaction coefficients for the simulated data from the null models.

[15] This seemed sufficient given the nature of the results presented below. In one case for Experiment 3, we went all the way to 10,000 times as discussed below.

Probability Model	1st 1000 simulations	2nd 1000 simulations
All generalizations evaluated together	0.831	0.796
One generalization evaluated at a time	0.966	0.945

Table 5: Proportion of interaction coefficients for data simulated from the null models that were further away from the mean of the interaction coefficients than the actual interaction observed in Experiment 2.

3.4 Discussion

In Experiment 1, there was a possible confound that learners were also keeping track of consonant sequence generalizations, which might have affected their responses to the NewStims. If segments themselves are representational primitives, then the super-additive effect observed for the NewStims in Experiment 1 could have been accounted for by multiple viewpoints. In Experiment 2, once the segment sequence confound was removed from the NewStims test stimuli by withholding the relevant consonant sequences during training, the proportion of 'Yes' responses to the NewStims was no more than an additive effect of the same for the OnlyVoicing and OnlyContinuancy stimuli (see Figure 4 and Table 4); that is, there was no super-additive effect. This suggests that learners do not keep track of more complex featural generalizations when simpler generalizations are available. Furthermore, a closer look at the cumulative proportion of 'Yes' responses for each type of test stimulus (Figure 5) revealed that the lack of a super-additive effect for the NewStims could not be due to insufficient statistical power; *i.e.*, this is not due to an insufficient number of participants in our experiment. Finally, the results also clearly establish that learners were indeed learning both simple generalizations, since the higher the 'Yes' responses to OnlyVoicing over Disharmony, the higher the responses to OnlyContinuancy over Disharmony (Figure 6).

There are three aspects of the data that deserve further consideration. First, the proportion of 'Yes' responses to OnlyVoicing over Disharmony, while consistent in direction with the results in Experiment 1, was just barely statistically significant. Second, the responses to the NewStims in Experiments 1 & 2, while visually different, were not directly comparable due to huge imbalances in the number of participants and differences in the types of training and test stimuli. Finally, given

that there can be subtle effects of the training data on potential generalizations, as discussed in Gerken and Knight (2015), it is important to establish that the results are not due to accidental patterns in the (randomized) training stimuli. We ran Experiment 3, which is in part a replication of Experiment 2, to address these concerns.

4 Experiment 3

4.1 Methods

4.1.1 Participants

51 English-speaking undergraduates at Michigan State University participated in this experiment for extra credit (Mean(Age)=20.1yrs, SD(Age)=3.6yrs; 39 Females & 12 Males). None of the participants were excluded due to non-learning.[16]

4.1.2 Materials

The design of the Experiment 3 was nearly identical to those of Experiments 1 & 2. Again, the possible vowels and consonants were the same, and the experiment also took about 10–15 minutes overall. The minor differences between Experiment 3 and Experiments 1 & 2 are discussed immediately below.

Training The training phase for Experiment 3 was identical to that of Experiment 2. As in Experiment 2, we withheld certain consonant sequences in the training phase of Experiment 3. The training stimuli were presented in the exact same manner as in Experiments 1 and 2.

[16] It is interesting to note that, unlike in Experiments 1 & 2, none of the participants hit ceiling for both the Disharmony and OldStims test items. It is unclear what caused this change in participant results. Since all instructions were the same and were provided through the experimental software, we are unaware of any systematic changes in pre-experiment instructions given to the participants. It is possible that the emphasis to some of the participants by one of the authors to focus on the training might have had an effect; however, it is not obvious how that has a bearing on the results.

Testing The testing phase in Experiment 3 was nearly identical to the testing phase in Experiment 2. However, there were 6 different types of testing stimuli (instead of 5), consisting of 10 items each for a total of 60 test items.

As in Experiments 1 & 2, the test items consisted of Disharmony, OnlyVoicing, OnlyContinuancy, and OldStims stimuli. Along with those 4 types, there were two other types of NewStims, corresponding to the NewStims of Experiments 1 & 2. Those new stimuli that consist of consonant sequences observed during training, as in Experiment 1, are labeled NewWordStims. To reiterate, these are novel stimuli because the vowels are different. For example, a participant might have heard [fusi] during training but not [fisa]; [fisa] could have therefore been a NewWordStim for this participant in the test phase. So while the consonant sequence is not new, the word itself is new to the participant. Furthermore, we did still withhold random consonant sequences from participants on an individual basis in the training phase, just like in Experiment 2. The withheld consonant sequences were used for what we call the NewConsStims testing stimuli. For example, if a participant never heard [bVdV] and [dVbV] during training, then any words of this form could have made up the NewConsStims for this participant in the testing phase.

4.2 Predictions

Since the experiment was done to (a) confirm the findings of the previous two experiments; (b) allow for a more direct comparison of the 'Yes' responses to the two types of NewStims in Experiments 1 & 2; and (c) to ensure that the effect observed for OnlyVoicing stimuli in Experiment 2 was replicable, the predictions of all the viewpoints presented earlier are effectively the same as discussed in Sections 2.2 (see Figure 2). We furthermore predict NewWordStims to be rated more highly than NewConsStims, because NewWordStims also conform to any segment-based generalizations a learner might have formed during training (in addition to the feature-based generalizations), whereas the NewConsStims do not.

4.3 Results

A visual inspection of the mean 'Yes' responses suggests that the five types of test stimuli of main interest (OnlyVoicing, OnlyContinuancy, NewConsStims, NewWordStims and OldStims) are more acceptable to participants than the Disharmony stimuli (Figure 7). Three further observations can be made. First, the OnlyVoicing stimuli are clearly more acceptable than Disharmony stimuli, thereby suggesting that the marginally significant results in Experiment 2 were not due to chance variation. Second, as in Experiment 1, the 'Yes' responses for the NewWordStims appear to be the result of a super-additive effect of the 'Yes' responses to the OnlyVoicing and OnlyContinuancy stimuli, over and above the Disharmony stimuli. Finally, as in Experiment 2, the 'Yes' responses for the NewConsStims appear to be just an additive effect (if anything, *sub-additive* effect) of the 'Yes' responses to the OnlyVoicing and OnlyContinuancy stimuli, over and above the Disharmony stimuli.

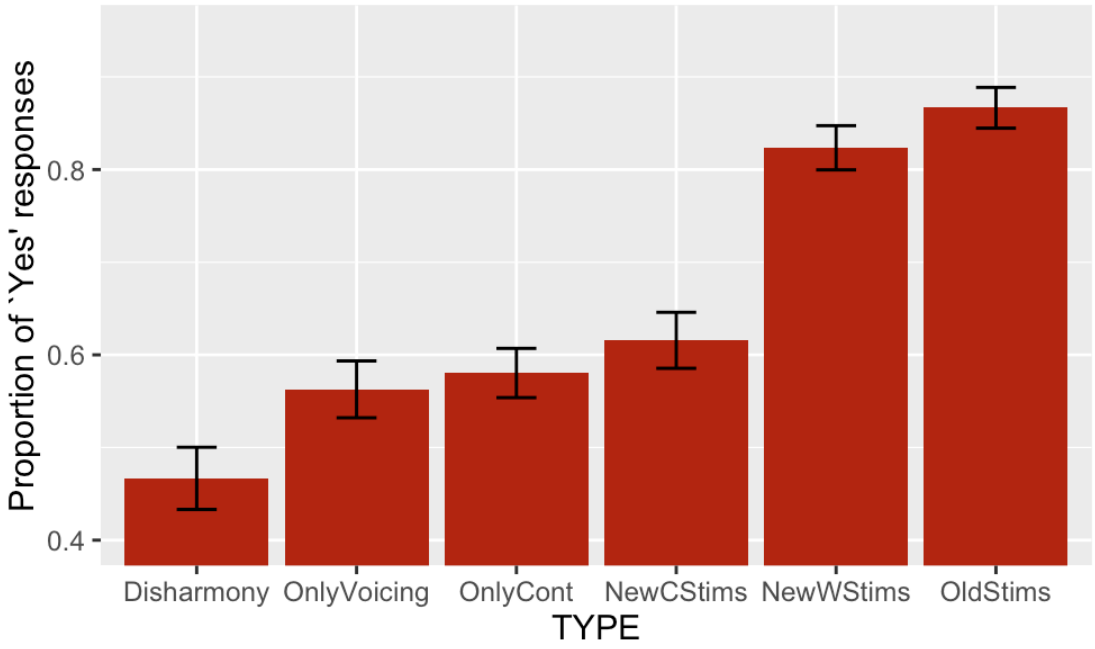

Figure 7: Proportion of 'Yes' responses to test stimuli in Experiment 3 (error bars represent standard errors; OnlyCont = OnlyContinuancy; NewCStims = NewConsStims; NewWStims = NewWordStims).

As in Experiments 1 & 2, to find out if the responses to the NewConsStims were more than

an additive effect of the OnlyVoicing and OnlyContinuancy responses (over and above the Disharmony stimuli), we coded the OnlyVoicing stimuli as *Voicing*, the OnlyContinuancy stimuli as *Continuancy*, the NewConsStims as both *Voicing* and *Continuancy*, and the Disharmony stimuli as neither *Voicing* nor *Continuancy*. And again, we attempted to fit a mixed-effects logistic regression model (following the procedure discussed in Section 2.3). The random effects structure, as in Experiments 1 & 2, included a varying intercept for subjects and items. As with the results of Experiment 2, the best model was one with two simple main effects (Table 6).[17] Therefore, the results of Experiment 3 replicate those of Experiment 2.

Fixed Effect	Estimate	z-value	Pr($>z$)	
(Intercept)	–0.0689	–0.538	=0.295	
Voicing	0.3022	2.948	<0.01	**
Continuancy	0.3737	3.646	<0.001	***

Table 6: Best fitting logistic mixed-effects model for Experiment 3

Next, to compare if the proportion of 'Yes' responses to NewWordStims was higher than that for NewConsStims, we fit a logistic mixed-effects model with the data subsetted to only those two types of test stimuli, and the responses to NewConsStims as the baseline. Therefore, the independent variable of *Type* has only two levels (NewConsStims, NewWordStims). The random effects structure was one with a varying intercept for both subjects and items. The model with the independent factor for *Type* was the best model for the above random effects structure (Table 7). The model clearly supports the earlier visual inspection in suggesting that there was indeed a higher proportion of 'Yes' responses to NewWordStims compared to NewConsStims.

Fixed Effect	Estimate	z-value	Pr($>z$)	
(Intercept)	0.5300	3.523	<0.001	***
NewWordStims	1.2169	7.335	<0.0001	***

Table 7: Logistic mixed-effects model comparing NewConsStims and NewWordStims

Finally, as with Experiment 2, we also wanted to take a closer look at the data to see if partici-

[17] The model with the interaction term for *Voicing* and *Continuancy* was not significantly better than the best model. Furthermore, the interaction term was not significant in that model ($\hat{\beta} = -0.29, p = 0.16$).

pants were indeed learning both simple generalizations (*i.e.*, [α voice] and [β cont]). As can be seen in Figure 8, replicating the results of Experiment 2, there is a positive correlation between the preference for OnlyContinuancy and the preference for OnlyVoicing ($\hat{\beta} = 0.372, p = 0.001$). Therefore, there is again no tradeoff between learning the two generalizations for the learners. This suggests, in line with Experiment 2, that learners who learned Voicing harmony also learned Continuancy harmony, thereby clearly showing that participants are able to learn both simple generalizations simultaneously.

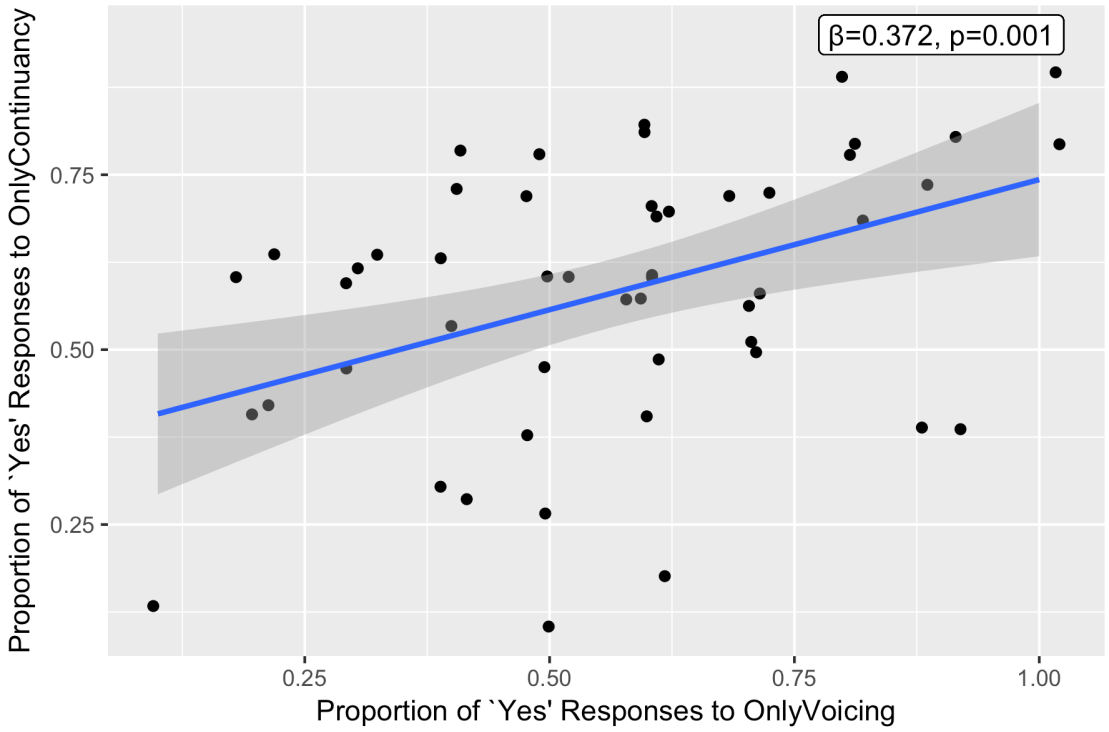

Figure 8: Correlation between increase in 'Yes' responses in Experiment 3 to OnlyVoicing and OnlyContinuancy (note: jitter has been added to the plot to reveal overlapping data points).

4.3.1 Is there any evidence that the complex generalization was learned?

As with Experiment 2, to probe whether the more complex generalization was also learned along with the simple generalizations, we ran a Monte Carlo simulation. Here, too, we ran the simulations with both probability models described in Section 3.3.1. (Again, see Supplementary Materials A for an R script that reproduces these simulations.)

Below, we present the results based on two sets of 1,000 simulations of the null models for Experiment 3. Table 8 shows the results of the simulations. The second set of simulations for the second probability model (that assumes averaging over the probabilities of all generalizations is appropriate) is based on 10,000 replications (this value is italicized in the table). This was done because the observed proportion in the first 1,000 replications was very close to the 0.05 threshold, so we thought it prudent to get a larger set of interaction coefficients to compare with.

Probability Model	1^{st} simulation	2^{nd} simulation
All generalizations evaluated together	0.2150	0.2360
One generalization evaluated at a time	0.0930	*0.0812*[18]

Table 8: Proportion of interaction coefficients for data simulated from the null models that were further away from the mean of the interaction coefficients than the actual interaction observed in Experiment 3.

As with Experiment 2, there is no clear evidence of an interaction effect beyond the simple model. There is at best a marginally significant effect, under the second probability model (that assumes that the learner uses only one generalization at any one time during evaluation). But, if one were to take the gamut of results, it is clear from Experiments 2 & 3, that our results support the null models (*i.e.*, the models with only the simple generalizations).[19]

4.4 Discussion

In Experiment 3, we were able to replicate all the important aspects of the results in Experiment 2. First, there is a decrement in the preference for NewConsStims compared to NewWordStims,

[18] This is based on 10,000 replications. Note, the proportion of values below the observed interaction coefficient would be half the proportions presented in the table; however, these are effectively one-tailed p-values, which we think are inappropriate given the direction of the interaction parameter in the experiment was not predicted to be less than the mean of the interaction parameters for the simulated data *a priori*.

[19] Furthermore, the marginally significant effects are only for the probability model where the generalizations can be used *individually* and *mutually exclusively* of one another while making an acceptability judgments, despite there being multiple generalizations present as part of the grammar. This behavior is not possible to account for with phonological grammars that have parallel architectures such as Optimality Theory and Harmonic Grammar.

paralleling the decrement in preference for the NewStims in Experiment 2 compared to those in Experiment 1. This suggests that there are other generalizations (possibly, consonant sequence generalizations) that allowed the participants to rate the NewWordStims in Experiment 3 and the NewStims in Experiment 1 so highly.

Second, the preference for the OnlyVoicing stimuli over Disharmony stimuli was just barely statistically significant in Experiment 2, so it was important to see that the effect replicated; it was indeed replicated in Experiment 3.

Finally, and most crucially, as in Experiment 2, the preference for the NewConsStims stimuli was no more than an additive effect of the preference for OnlyVoicing and OnlyContinuancy stimuli. Therefore, this reinforces the finding that the learners in our experiments were simply not learning the more complex generalization when simpler generalizations were available.

5 General discussion and conclusion

We presented the results of three ALL experiments that probed the question of what learners do when faced with data that is consistent with multiple competing phonotactic generalizations. As mentioned earlier, it is more insightful to break down the question into two sub-questions: (a) do learners learn just a single generalization that is consistent with a specific segmental co-occurrence pattern, or do they learn multiple (even all) possible generalizations; and (b) do learners learn the simplest generalization or the most specific (therefore, most complex) generalization? Our results suggest that learners are indeed able to keep track of multiple generalizations (Experiments 1, 2, & 3). However, this does not mean they keep track of *all* available generalizations. While the results of Experiment 1 appeared to suggest that learners could be learning more complex generalizations, when the confounding possibility of using consonant sequence patterns was removed from the relevant test items (NewStims in Experiment 2, and NewConsStims in Experiment 3), learners showed no evidence of learning the more complex featural generalization; instead, participants were only keeping track of the simplest generalizations, where simplest is defined as the use of the

fewest representational primitives needed to state the generalization.

As briefly mentioned before in Section 1, in the interest of expository convenience, we have conflated the terms 'most complex' and 'most specific', where the former refers to the intensional description, while the latter refers to the extension set. Crucially, the learners in our experiments are able to learn both simple featural generalizations and segment based generalization, but do not seem to be able to learn the complex featural generalization. The fact that learners learn multiple simple featural generalizations but are unable to learn the complex featural generalization suggests that *simplicity* is an important notion in understanding learnable/unlearnable patterns. In contrast, the fact that the learner is able to learn both the most specific segmental generalizations and the least specific single feature generalizations, but is unable to learn the complex feature generalization (which is in between on the specificity scale) suggests that *specificity* is not a useful notion in understanding learnable/unlearnable patterns.[20] To reiterate, our experiments suggest that the notion of *simplicity* is of relevance to the learner, but not the notion of *specificity*.

It is important to note that the results cannot be accounted for by just saying simpler generalizations are easier to learn (Cristià and Seidl 2008; Kuo 2009; Pycha et al. 2003; Saffran and Thiessen 2003). Such a statement is insufficient to account for the results as there was ample training data[21] provided to participants to learn the more complex generalization (compared to previous research that showed that complex generalizations appear to be learned, when there is no competition with simpler generalizations). Therefore, even by this view, the complex generalization could still have been learned, albeit with less weight attached to it. If so, there should have been a super-additive

[20] To the extent that more specific grammars are favored as a matter of fact by the math behind Bayesian inference (cf. the Size Principle; Linzen and O'Donnell 2015; Tenenbaum and Griffiths 2001; Xu and Tenenbaum 2007), this may tell against some of the predictions of Bayesian models of learning. This may be due to the fact Bayesian models are not necessarily incremental/algorithmic models of learning, but rather computational-level models; this is worth further investigating.

[21] We say the training data was ample compared to other ALL experiments. Of course, to defend any particular hypothesis, one could always say that it is an insufficient number of training items and insufficient number of training segments in the items. In which case, we think the onus is on such researchers to specify what constitutes sufficient training data to falsify the hypothesis. For example, there could be a strong prior bias against a grammar with the conjoined feature rule, and we would only see evidence for the learning of this conjoined feature rule if participants received more data. This may be true, but this could always be true in the absence of a clear statement of what one thinks the magnitude of the prior bias is. In particular, in the absence of an idea of what the magnitude of the bias against a conjoined feature rule might be, we don't think it's a fruitful discussion to have.

effect observed in participants' preference for the relevant NewStims (or NewConsStims) stimuli in Experiments 2 & 3. Furthermore, some have suggested that the learning bias for simpler generalizations stems from what amounts to a sampling bias (Pierrehumbert 2001a,b). Pierrehumbert suggested that one reason simpler generalizations are learned more than more complex ones is possibly due to more complex generalizations requiring a more specific set of data to be confirmed, and random sampling might not allow the learner to experience that particular set of data. However, in our experiments, the amount of training data that supported the more complex generalization was equal to the amount of training data that supported the simpler generalizations; therefore, it is clear the bias observed cannot be reduced to a sampling bias in the input data.

There are four other issues that we wish to touch upon. First, how do we square our results with those of Linzen and Gallagher (2014, 2017) and Gerken (2006) who argue that their results suggest that learners might be learning more complex (or specific) generalizations? As a reviewer points out (and in our opinion), it is reasonable to reinterpret Gerken's (2006) results as showing that learners need stimulus variation to infer a more abstract generalization. For, if the learner were simply holding on to the subset grammar, they could still have memorized all the possible final syllables. Furthermore, what needs to be pointed out in each of the above papers is that the specific or more complex generalizations involved mixing representational primitives. For example, in the case of Linzen and Gallagher (2014, 2017), their specific generalization involved segments, while the simpler one involved features. However, the specific one is only more complex under the assumption that segments are not representational primitives by themselves and are nothing more than a collection of features. If this assumption is wrong, then equating specificity to the notion of complexity is not correct for their experiments.[22] As we showed in our experiments, when the possibility of segment sequence generalizations was removed (Experiments 2 & 3), there was no more evidence for the claim that learners were keeping track of more complex (featural) generalizations. Similarly, Gerken (2006) also compared a simple syllabic generalization (AAB) versus one that involved both syllables and segments (AAdi). Given that the putative complex general-

[22] We do not attribute this 'error' to Linzen and Gallagher (2014, 2017). It is important to reiterate here that their results were interpreted as possible evidence for the *Proportional to Simplicity* viewpoint by the authors of this article.

ization involved both syllables and segments, it is possible that the more complex generalization is actually decomposable into simpler generalizations involving syllables and segments separately. What we wish to primarily highlight from this discussion is that in order to test the learnability of simple vs. complex generalizations, the class of representational primitives used in the two types of generalizations needs to be kept constant, as was done in our Experiments 2 & 3. Otherwise, it is difficult to draw firm conclusions from the results, whatever they may be.

Second, how do we square the results with those of Finley (2011, 2012), Lai (2015), and McMullin (2016) whose results suggests that learners appear to be unwillingly to accept a seemingly more general non-local pattern across both vowels and consonants ("second order non-local") when trained on a transvocalic ("first order local") pattern but are willing to accept a transvocalic pattern, when trained on a more general non-local pattern?[23] In our opinion, these results are actually unclear with respect to the issue of simplicity. The crux of the argument in such experiments is contingent on comparing against chance (0.5) and inferring that a generalization has not been used in novel contexts if the acceptability proportion is around 0.5; however, it is not clear that a proportion of 0.5 is the appropriate representation of chance (as the stimuli have other patterns in them that are consistent with the training data); instead the results should be seen in relative terms in our opinion. That is, when trained on transvocalic patterns, participants accept transvocalic patterns *more than* the more general pattern, but when trained on the more general (trans-segmental) pattern, there is no such clear difference. If the results are seen in this light, we can in fact reinterpret the results. Following Heinz (2010), if a learner is equipped with the ability to learn both n-gram generalizations (up to a suitable "n"; see Cowan (2010) for an argument that the n≈4 for short-term memory generally) and *separate* precedence/piecewise generalizations which make no reference to locality, then, when presented with training stimuli with transvocalic harmony, learners could represent them with both an n-gram generalization and a precedence generalization, and as a consequence, during testing, the transvocalic stimuli are going to get an additive effect on acceptance from both the types of generalizations, while the more general (trans-segmental) generalizations

[23] It is worth remembering that in the second order non-local case, the patterns were across VCV contexts (so, _VCV_), while in the first order local case, the patterns were across a single V (so, _V_).

are acceptable only due to the precedence grammar. In effect, we expect the latter to be more acceptable than the former. In contrast, when presented with the trans-segmental pattern during training, the learner can only learn the precedence generalization; as a consequence, during testing, they show no difference between the transvocalic and trans-segmental patterns.

Third, the fact that learners keep track of *only* the simplest generalization(s) consistent with the data in the face of ambiguity suggests that there is a certain structure to the search space of possible generalizations, as touched upon by Chomsky and Halle (1968) and Hayes and Wilson (2008); this structure, if present, dramatically decreases the computational challenge faced by the learner.[24] Such a view leads to a slight reinterpretation of previous ALL results that suggest that simpler generalizations are easier to learn than more complex generalizations (Cristià and Seidl 2008; Kuo 2009; Pycha et al. 2003; Saffran and Thiessen 2003). More specifically, the results presented in this article suggest that the reason simpler generalizations appear to be learned better in previous ALL experiments is that learners attempt to learn simpler generalizations first, and, only in the absence of viable simpler generalizations, do they attempt to learn more complex ones.

Fourth, a more speculative possibility, one that takes a substantial inductive leap from our results, is that learners are only able to keep track of simple phonotactic generalizations, by which we mean generalizations that involve the precedence relationships between at most a single pair of features, segments, syllables, *etc.*. Therefore, the issue of complex vs. simple learned generalizations itself would vanish, as the learner simply cannot keep track of complex generalizations (of the relevant type).[25] For example, learners might be able to keep track of precedence relationships such as [feature1]…[feature2], or [segment1]…[segment2], *etc.*, but not precedence relationships involving more than that, such as [feature1,feature3]…[feature2,feature3], or in the case of segments [segment1]…[segment2]…[segment3]. Such a possibility would automatically explain why our experiments found no evidence for learners learning the complex generalization. It would further suggest that the reason previous experiments seemed to show learning of a complex generalization

[24]However, this is not to say that the learning task in any sense has become trivial or 'easy'.

[25]Note, a similar sentiment that simple (and even categorical) models of phonotactics can account for the extant data on word-acceptability judgments was discussed by Gorman (2013).

was just because they did not—or could not, given their design—test whether their results were merely the additive result of multiple simple generalizations. As mentioned above, this possibility requires a big inductive leap from the results presented in this article and should at this point be seen as a speculation that is, at best, worthy of future analytical/experimental consideration.

In conclusion, we would like to reiterate the primary findings in the article. In the face of data that is ambiguous between many phonotactic sequence generalizations, learners keep track of multiple generalizations, as long as they are all the simplest possible generalizations.

References

Albright, Adam (2009). "Feature-based generalisation as a source of gradient acceptability." *Phonology* 26 (01), pp. 9–41. ISSN: 1469-8188. doi: 10.1017/S0952675709001705.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal." *Journal of Memory and Language* 68, pp. 255–278.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67.1, pp. 1–48. doi: 10.18637/jss.v067.i01.

Becker, Michael, Nihan Ketrez, and Andrew Nevins (2011). "The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations." *Language* 87.1, pp. 84–125.

Bergelson, Elika and William J. Idsardi (Apr. 2009). "Structural Biases in Phonology: Infant and Adult Evidence from Artificial Language Learning." *BUCLD 33: Proceedings of the 33rd Annual Boston University Conference on Language Development* 1. Ed. by Jane Chandlee, Michelle Franchini, Sandy Lord, and Gudrun-Marion Rheiner, pp. 85–96.

Berwick, Robert C. (Aug. 1985). *The Acquisition of Syntactic Knowledge*. Vol. 16. Artificial Intelligence. Cambridge (MA) and London: The MIT Press.

Bolker, Benjamin M. (2014). *Random- and fixed-effects structure in linear-mixed models*. Cross Validated (StackExchange). url: http://stats.stackexchange.com/questions/130714/random-and-fixed-effects-structure-in-linear-mixed-models.

Chambers, Kyle E., Kristine H. Onishi, and Cynthia Fisher (2003). "Infants learn phonotactic regularities from brief auditory experience." *Cognition* 87.2, B69–B77. doi: 10.1016/s0010-0277(02)00233-0.

Chambers, Kyle E., Kristine H. Onishi, and Cynthia Fisher (2012). "Representations for phonotactic learning in infancy." *Language Learning and Development* 7.4, pp. 287–308.

Chomsky, Noam and Morris Halle (1968). *The Sound Pattern of English*. New York, Evanston, and London: Harper and Row.

Cowan, Nelson (2010). "The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why?" *Current Directions in Psychological Science* 19.1. PMID: 20445769, pp. 51–57. doi: 10.1177/0963721409359277. eprint: https://doi.org/10.1177/0963721409359277.

Cristià, Alejandrina, Jeff Mielke, Robert Daland, and Sharon Peperkamp (2013). "Similarity in the generalization of implicitly learned sound patterns." *Laboratory Phonology* 4.2, p. 259. issn: 18686346. doi: 10.1515/lp-2013-0010.

Cristià, Alejandrina and Amanda Seidl (2008). "Is Infants' Learning of Sound Patterns Constrained by Phonological Features?" *Language Learning and Development* 4.3, pp. 203–227. doi: 10.1080/15475440802143109. eprint: http://dx.doi.org/10.1080/15475440802143109.

Cristià, Alejandrina, Amanda Seidl, and LouAnn Gerken (2011). "Learning classes of sounds in infancy." *University of Pennsylvania Working Papers in Linguistics* 17.1, pp. 69–76.

Culbertson, Jennifer (2012). "Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning." *Language and Linguistics Compass* 6.5, pp. 310–329. issn: 1749-818X. doi: 10.1002/lnc3.338.

Dell, François (1981). "On the Learnability of Optional Phonological Rules." *Linguistic Inquiry* 12.1, pp. 31–37.

Dupoux, E., K. Kakehi, Y. Hirose, C. Pallier, and J. Mehler (1999). "Epenthetic vowels in Japanese: A perceptual illusion?" *Journal of Experimental Psychology-human Perception and Performance* 25.6, pp. 1568–1578.

Eberhardt, Frederick and David Danks (May 2011). "Confirmation in the Cognitive Sciences: The Problematic Case of Bayesian Models." *Minds and Machines* 21.3, pp. 389–410.

Field, Andy P. (2013). *Discovering statistics using IBM SPSS Statistics*. London, Thousand Oaks, New Delhi: SAGE publications.

Finley, Sara (July 2011). "The Privileged Status of Locality in Consonant Harmony." *Journal of Memory and Language* 65.1, pp. 74–83. doi: 10.1016/j.jml.2011.02.006.

Finley, Sara (May 2012). "Testing the Limits of Long-Distance Learning: Learning Beyond a Three-Segment Window." *Cognitive Science* 36.4, pp. 740–756. doi: 10.1111/j.1551-6709.2011.01227.x.

Finley, Sara and William Badecker (2009). "Artificial language learning and feature-based generalization." *Journal of Memory and Language* 61.3, pp. 423–437. doi: 10.1016/j.jml.2009.05.002.

Folia, Vasiliki, Julia Uddén, Meinou De Vries, Christian Forkstam, and Karl Magnus Petersson (2010). "Artificial Language Learning in Adults and Children." *Language Learning* 60, pp. 188–220. issn: 1467-9922. doi: 10.1111/j.1467-9922.2010.00606.x.

Friederici, Angela D. and Jeanine M. I. Wessels (1993). "Phonotactic knowledge of word boundaries and its use in infant speech perception." *Perception & Psychophysics* 54.3, pp. 287–295. issn: 1532-5962. doi: 10.3758/BF03205263.

Gerken, LouAnn (2006). "Decisions, decisions: infant language learning when multiple generalizations are possible." *Cognition* 98, B67–B74.

Gerken, LouAnn and Sara Knight (Oct. 2015). "Infants Generalize from Just (the Right) Four Words." *Cognition* 143, pp. 187–192. doi: 10.1016/j.cognition.2015.04.018.

Gorman, Kyle (2013). "Generative Phonotactics." PhD dissertation. University of Pennsylvania.

Hale, Mark and Charles Reiss (2003). "The Subset Principle in phonology: why the tabula can't be rasa." *Journal of Linguistics* 39.02, pp. 219–244.

Halle, Morris (1961). "On the Role of Simplicity in Linguistic Descriptions." In: *Proceedings of Symposia in Applied Mathematics: Structure of Language and its Mathematical Aspects*. Vol. 12. American Mathematical Society, pp. 89–94.

Hayes, Bruce and Colin Wilson (2008). "A Maximum Entropy Model of Phonotactics and Phonotactic Learning." 39.3, pp. 379–440. doi: 10.1162/ling.2008.39.3.379.

Heinz, Jeffrey (2010). "Learning Long-Distance Phonotactics." *Linguistic Inquiry* 41.4, pp. 623–661. issn: 00243892, 15309150.

Jaeger, T. Florian (2009). *Random effect: Should I stay or should I go?*

Jaeger, T. Florian (2011). *More on random slopes and what it means if your effect is not longer significant after the inclusion of random slopes*.

Jusczyk, P.W., A.D. Friederici, J.M.I. Wessels, V.Y. Svenkerud, and A.M. Jusczyk (1993). "Infants' Sensitivity to the Sound Patterns of Native Language Words." *Journal of Memory and Language* 32.3, pp. 402–420. issn: 0749-596X. doi: http://dx.doi.org/10.1006/jmla.1993.1022.

Kabak, B. and W. J. Idsardi (2007). "Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints?" *Language and Speech* 50, pp. 23–52.

Kazanina, Nina, Jeffrey S. Bowers, and William Idsardi (2017). "Phonemes: Lexical access and beyond." *Psychonomic Bulletin & Review*. issn: 1531-5320. doi: 10.3758/s13423-017-1362-0.

Kuo, Li-Jen (2009). "The Role of Natural Class Features in the Acquisition of Phonotactic Regularities." *Journal of Psycholinguistic Research* 38.2, pp. 129–150. issn: 1573-6555. doi: 10.1007/s10936-008-9090-2.

Lai, Regine (2015). "Learnable vs. Unlearnable Harmony Patterns." *Linguistic Inquiry* 46.3, pp. 425–451. issn: 1749-818X. doi: 10.1162/ling_a_00188.

Linzen, Tal and Gillian Gallagher (2014). "The Timecourse of Generalization in Phonotactic Learning." In: *Proceedings of the 2013 Meeting on Phonology*. Ed. by John Kingston, Claire Moore-Cantwell, Joe Pater, and Robert Staubs. Washington D.C.: Linguistic Society of America, pp. 1–12. doi: 10.3765/amp.v1i1.18.

Linzen, Tal and Gillian Gallagher (Oct. 2017). "Rapid Generalization in Phonotactic Learning." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8.1, pp. 1–32. doi: 10.5334/labphon.44.

Linzen, Tal and Timothy J. O'Donnell (2015). "A Model of Rapid Phonotactic Generalization." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1126–1131.

McMullin, Kevin (Feb. 2016). "Tier-Based Locality in Long-Distance Phonotactics: Learnability and Typology." PhD dissertation. Vancouver, Canada: University of British Columbia.

McQueen, James M. (1998). "Segmentation of Continuous Speech Using Phonotactics." *Journal of Memory and Language* 39.1, pp. 21–46. issn: 0749-596X. doi: http://dx.doi.org/10.1006/jmla.1998.2568.

Moreton, Elliott (2002). "Structural constraints in the perception of English stop-sonorant clusters." *Cognition* 84, pp. 55–71.

Moreton, Elliott (May 2008). "Analytic Bias and Phonological Typology." *Phonology* 25 (1), pp. 83–127. doi: 10.1017/S0952675708001413.

Moreton, Elliott and Joe Pater (Nov. 2012a). "Structure and Substance in Artificial-phonology Learning, Part I: Structure." *Language and Linguistics Compass* 6.11, pp. 686–701. doi: 10.1002/lnc3.363.

Moreton, Elliott and Joe Pater (Nov. 2012b). "Structure and Substance in Artificial-Phonology Learning, Part II: Substance." *Language and Linguistics Compass* 6.11, pp. 702–718. doi: 10.1002/lnc3.366.

Peirce, J. W., J. R. Gray, S. Simpson, M. R. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. Lindeløv (2019). "PsychoPy2: experiments in behavior made easy." *Behavior Research Methods*. doi: 10.3758/s13428-018-01193-y.

Pierrehumbert, Janet B. (2001a). "Probabilistic phonology: discrimination and robustness." In: *Probabilistic Linguistics*. Ed. by Rens Bod, Jennifer Hay, and Stefanie Jannedy. Cambridge, Massachusetts: MIT Press. Chap. 6, pp. 177–228. doi: 10.1080/01690960143000218.

Pierrehumbert, Janet B. (2001b). "Why phonological constraints are so coarse-grained." *Language and Cognitive Processes* 16.5-6, pp. 691–698. doi: 10.1080/01690960143000218.

Pycha, Anne, Pawel Nowak, Eurie Shin, and Ryan Shosted (2003). "Phonological rule-learing and its implications for a theory of vowel harmony." *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22)*, pp. 101–114.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Saffran, Jenny R. and Erik D. Thiessen (2003). "Pattern Induction by Infant Language Learners." *Developmental Psychology* 39.3, pp. 484–494. doi: 10.1037/0012-1649.39.3.484.

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

Tenenbaum, Joshua B. and T. L. Griffiths (2001). "Generalization, similarity, and Bayesian inference." *Behavioral and Brain Sciences* 24, pp. 629–640.

Wilson, Colin (2006). "Learning Phonology With Substantive Bias: An Experimental and Computational Study of Velar Palatalization." *Cognitive Science* 30.5, pp. 945–982. issn: 1551-6709. doi: 10.1207/s15516709cog0000_89.

Xu, Fei and Joshua B. Tenenbaum (Apr. 2007). "Word Learning as Bayesian Inference." *Psychological Review* 114 (2), pp. 245–272. doi: 10.1037/0033-295X.114.2.245.

A Code for Monte Carlo simulation

```
library(tidyverse)
library(lme4)

#Step 1: Functions for different probability models of acceptance ------

generalizations_evaluated_together <- function(data) {
  #Equations to be solved
  #ungrammatical under both generalizations
  #prob_yes_for_disharmonic =
  #      1 - (prob_no_if_ungrammatical * prob_no_if_ungrammatical)
  #log(prob_no_if_ungrammatical) = log(1-prob_yes_for_disharmonic) / 2
  prob_no_if_ungrammatical <-
    exp(log(1 - data$Mean_Yes[data$TYPE == "Disharmony"]) / 2)

  #grammatical under the voicing generalization
  #prob_yes_for_onlyVoice =
  #      1 - (prob_no_if_gramm_by_voice * prob_no_if_ungrammatical)
  #log(prob_no_if_gramm_by_voice) + log(prob_no_if_ungrammatical) =
  #      log(1 - prob_yes_for_onlyVoice)
  prob_no_if_gramm_by_voice <-
    exp(log(1 - data$Mean_Yes[data$TYPE == "OnlyVoicing"]) -
          log(prob_no_if_ungrammatical))

  #grammatical under the continuancy generalization
  #prob_yes_for_onlyCont =
  #      1 - (prob_no_if_ungrammatical * prob_no_if_gramm_by_cont)
```

```
#log(prob_no_if_gramm_by_cont) + log(prob_no_if_ungrammatical) =
#      log(1 - prob_yes_for_onlyCont)
prob_no_if_gramm_by_cont <-
  exp(log(1 - data$Mean_Yes[data$TYPE == "OnlyCont"]) -
        log(prob_no_if_ungrammatical))

#Get the yes probabilities
prob_yes_if_ungrammatical <- 1 - prob_no_if_ungrammatical
prob_yes_if_gramm_by_voice <- 1 - prob_no_if_gramm_by_voice
prob_yes_if_gramm_by_cont <- 1 - prob_no_if_gramm_by_cont

#Generate probabilities from above model
#ungrammatical under both generalizations
prob_yes_for_disharmonic <- 1 - (prob_no_if_ungrammatical *
                                   prob_no_if_ungrammatical)
#grammatical under the voicing generalization
prob_yes_for_onlyVoice <- 1 - (prob_no_if_ungrammatical *
                                 prob_no_if_gramm_by_voice)
#grammatical under the continuancy generalization
prob_yes_for_onlyCont <- 1 - (prob_no_if_ungrammatical *
                                prob_no_if_gramm_by_cont)
#grammatical under both simple generalizations
prob_yes_for_newStim <- 1 - (prob_no_if_gramm_by_voice *
                               prob_no_if_gramm_by_cont)

#Return the probabilities that a participant says a word is
#acceptable, depending on the nature of the word
```

```
  data.frame(prob_yes_for_disharmonic=prob_yes_for_disharmonic,
             prob_yes_for_onlyVoice=prob_yes_for_onlyVoice,
             prob_yes_for_onlyCont=prob_yes_for_onlyCont,
             prob_yes_for_newStim=prob_yes_for_newStim)
}

generalizations_evaluated_one_at_a_time <- function(data) {
  #Equations to be solved
  #ungrammatical under both generalizations
  #prob_yes_for_disharmonic =
  #      ((1 - prob_no_if_ungrammatical) +
  #       (1 - prob_no_if_ungrammatical)) / 2
  prob_no_if_ungrammatical <-
    (1 - data$Mean_Yes[data$TYPE == "Disharmony"])

  #grammatical under the voicing generalization
  #prob_yes_for_onlyVoice =
  #      ((1 - prob_no_if_gramm_by_voice) +
  #       (1 - prob_no_if_ungrammatical)) / 2
  #2 * prob_yes_for_onlyVoice =
  #      2 - prob_no_if_gramm_by_voice - prob_no_if_ungrammatical
  #prob_no_if_gramm_by_voice =
  #      2 - 2 * prob_yes_for_onlyVoice - prob_no_if_ungrammatical
  prob_no_if_gramm_by_voice <- 2 -
    2 * data$Mean_Yes[data$TYPE == "OnlyVoicing"] -
    prob_no_if_ungrammatical
```

```
#grammatical under the continuancy generalization
#prob_yes_for_onlyCont =
#      ((1 - prob_no_if_gramm_by_cont) +
#       (1 - prob_no_if_ungrammatical)) / 2
#2 * prob_yes_for_onlyCont =
#      2 - prob_no_if_gramm_by_cont - prob_no_if_ungrammatical
#prob_no_if_gramm_by_cont =
#      2 - 2 * prob_yes_for_onlyCont - prob_no_if_ungrammatical
prob_no_if_gramm_by_cont <- 2 -
  2 * data$Mean_Yes[data$TYPE == "OnlyCont"] -
  prob_no_if_ungrammatical

#Get the yes probabilities
prob_yes_if_ungrammatical <- 1 - prob_no_if_ungrammatical
prob_yes_if_gramm_by_voice <- 1 - prob_no_if_gramm_by_voice
prob_yes_if_gramm_by_cont <- 1 - prob_no_if_gramm_by_cont

#Generating probabilities from above model
#ungrammatical under both generalizations
prob_yes_for_disharmonic <- ((1 - prob_no_if_ungrammatical) +
                               (1 - prob_no_if_ungrammatical)) / 2
#grammatical under the voicing generalization
prob_yes_for_onlyVoice <- ((1 - prob_no_if_gramm_by_voice) +
                             (1 - prob_no_if_ungrammatical)) / 2
#grammatical under the continuancy generalization
prob_yes_for_onlyCont <- ((1 - prob_no_if_gramm_by_cont) +
                            (1 - prob_no_if_ungrammatical)) / 2
```

```
  #grammatical under both simple generalizations
  prob_yes_for_newStim <- ((1 - prob_no_if_gramm_by_voice) +
                             (1 - prob_no_if_gramm_by_cont)) / 2

  #Return the probabilities that a participant says a word is
  #acceptable, depending on the nature of the word
  data.frame(prob_yes_for_disharmonic=prob_yes_for_disharmonic,
             prob_yes_for_onlyVoice=prob_yes_for_onlyVoice,
             prob_yes_for_onlyCont=prob_yes_for_onlyCont,
             prob_yes_for_newStim=prob_yes_for_newStim)
}

#Step 2: Fit probabilities based on models to observed data ------------

exp2_data_summarized <-
  data.frame(TYPE = c("Disharmony", "OnlyVoicing", "OnlyCont",
                      "NewStims", "OldStims"),
             Mean_Yes = c(0.532, 0.57, 0.614, 0.653, 0.898),
             SE_Yes = c(0.0253, 0.0269, 0.025, 0.0281, 0.0140))

exp3_data_summarized <-
  data.frame(TYPE = c("Disharmony", "OnlyVoicing", "OnlyCont",
                      "NewCStims", "NewWStims", "OldStims"),
             Mean_Yes = c(0.467, 0.563, 0.58, 0.616, 0.824, 0.867),
             SE_Yes = c(0.0336, 0.0307, 0.0266, 0.0302, 0.0239, 0.022))

exp2_evaluated_together_predictions <-
```

```
  generalizations_evaluated_together(exp2_data_summarized)

exp2_evaluated_one_at_a_time_predictions <-
  generalizations_evaluated_one_at_a_time(exp2_data_summarized)

exp3_evaluated_together_predictions <-
  generalizations_evaluated_together(exp3_data_summarized)

exp3_evaluated_one_at_a_time_predictions <-
  generalizations_evaluated_one_at_a_time(exp3_data_summarized)

#Step 3: Simulate data and fit logisitic regression to results ---------

#Function for simulating judgment data
simulate_judgments <- function(num_participants,
                               num_items_per_type,
                               probability_model_predictions,
                               conditions) {
  TYPE = rep(conditions, num_items_per_type)
  judgments = data.frame(participant = rep(c(1:num_participants),
                                           each = num_items_per_type *
                                             length(conditions)),
                         TYPE = rep(TYPE, each = num_participants))

  #Simulate acceptability responses
  judgments <- judgments %>%
    #Ensure each row gets a different random number
```

```
    rowwise() %>%
    mutate(Random_Number = runif(1)) %>%
    ungroup() %>%
    mutate(response = case_when(
      TYPE == conditions[1] ~ Random_Number <=
        probability_model_predictions$prob_yes_for_disharmonic,
      TYPE == conditions[2] ~ Random_Number <=
        probability_model_predictions$prob_yes_for_onlyVoice,
      TYPE == conditions[3] ~ Random_Number <=
        probability_model_predictions$prob_yes_for_onlyCont,
      TYPE == conditions[4] ~ Random_Number <=
        probability_model_predictions$prob_yes_for_newStim,
      TRUE ~ NA)) %>%
    mutate(response = as.numeric(response))

  judgments
}

#Function for running simulations
run_simulations <- function(num_simulations,
                            seed,
                            num_participants,
                            num_items_per_type,
                            probability_model_predictions,
                            conditions) {
  AllBetaValuesIfNullTrue=NULL
  for(i in 1:num_simulations) {
```

```
if(!is.null(seed)) {
  set.seed(seed * i + i)
}
message("Running simulation number ", i)
sim_data <- simulate_judgments(num_participants,
                               num_items_per_type,
                               probability_model_predictions,
                               conditions) %>%
  #Add coding for model
  mutate(Voicing = case_when(
    TYPE == "NewStims" | TYPE == "NewCStims" |
      TYPE == "OnlyVoicing" ~ 1,
    TRUE ~ 0),
    Stopping = case_when(
      TYPE == "NewStims" | TYPE == "NewCStims" |
        TYPE == "OnlyCont" ~ 1,
      TRUE ~ 0),
    Voicing = factor(Voicing),
    Stopping = factor(Stopping))

message("Fitting model for simulation number ", i)
model <- glmer(data = sim_data,
               response ~ Voicing * Stopping + (1|participant),
               family = binomial())

#Get coefficients from model
AllBetaValuesIfNullTrue <- rbind(AllBetaValuesIfNullTrue,
```

```
                                    coef(model)$participant[1, ])
  }
  AllBetaValuesIfNullTrue
}

#Run two different experiment 2 simulations
#Note that seed can be set to NULL or some other number for different
#simulations
exp2_evaluated_together_simulated_betas_1 <-
  run_simulations(num_simulations = 1000,
                  seed = 1234,
                  num_participants = 63,
                  num_items_per_type = 12,
                  probability_model_predictions =
                    exp2_evaluated_together_predictions,
                  conditions = c("Disharmony", "OnlyVoicing",
                                 "OnlyCont", "NewStims"))

exp2_evaluated_one_at_a_time_simulated_betas_1 <-
  run_simulations(num_simulations = 1000,
                  seed = 1234,
                  num_participants = 63,
                  num_items_per_type = 12,
                  probability_model_predictions =
                    exp2_evaluated_one_at_a_time_predictions,
                  conditions = c("Disharmony", "OnlyVoicing",
                                 "OnlyCont", "NewStims"))
```

```
exp2_evaluated_together_simulated_betas_2 <-
  run_simulations(num_simulations = 1000,
                  seed = 5678,
                  num_participants = 63,
                  num_items_per_type = 12,
                  probability_model_predictions =
                    exp2_evaluated_together_predictions,
                  conditions = c("Disharmony", "OnlyVoicing",
                                 "OnlyCont", "NewStims"))

exp2_evaluated_one_at_a_time_simulated_betas_2 <-
  run_simulations(num_simulations = 1000,
                  seed = 5678,
                  num_participants = 63,
                  num_items_per_type = 12,
                  probability_model_predictions =
                    exp2_evaluated_one_at_a_time_predictions,
                  conditions = c("Disharmony", "OnlyVoicing",
                                 "OnlyCont", "NewStims"))

#Run two different experiment 3 simulations
#Note that seed can be set to NULL or some other number for different
#simulations
exp3_evaluated_together_simulated_betas_1 <-
  run_simulations(num_simulations = 1000,
                  seed = 1234,
```

```
                    num_participants = 51,
                    num_items_per_type = 10,
                    probability_model_predictions =
                      exp3_evaluated_together_predictions,
                    conditions = c("Disharmony", "OnlyVoicing",
                                   "OnlyCont", "NewCStims"))

exp3_evaluated_one_at_a_time_simulated_betas_1 <-
  run_simulations(num_simulations = 1000,
                    seed = 1234,
                    num_participants = 51,
                    num_items_per_type = 10,
                    probability_model_predictions =
                      exp3_evaluated_one_at_a_time_predictions,
                    conditions = c("Disharmony", "OnlyVoicing",
                                   "OnlyCont", "NewCStims"))

exp3_evaluated_together_simulated_betas_2 <-
  run_simulations(num_simulations = 1000,
                    seed = 5678,
                    num_participants = 51,
                    num_items_per_type = 10,
                    probability_model_predictions =
                      exp3_evaluated_together_predictions,
                    conditions = c("Disharmony", "OnlyVoicing",
                                   "OnlyCont", "NewCStims"))
```

```
exp3_evaluated_one_at_a_time_simulated_betas_2 <-
  run_simulations(num_simulations = 10000,
                  seed = 5678,
                  num_participants = 51,
                  num_items_per_type = 10,
                  probability_model_predictions =
                    exp3_evaluated_one_at_a_time_predictions,
                  conditions = c("Disharmony", "OnlyVoicing",
                                 "OnlyCont", "NewCStims"))

#Step 4: Get interaction coefficient from actual data ------------------

#These are the interaction terms from models of the actual data
exp2_interaction_term <- 0.0147592
exp3_interaction_term <- -0.2890619

#Step 5: Actual interaction term further away from simulated mean? -----

get_prop_further <- function(sim_betas,
                             actual_interaction,
                             alternative = "two.sided") {
  #What proportion of interaction terms for the data based on the null
  #models are further away than the actual experimental results?
  if(alternative == "one.sided") {
    prop <- sum(sim_betas[, c("Voicing1:Stopping1")] >
                  actual_interaction) / NROW(sim_betas)
  }
```

```
  #What proportion of interaction terms from the simulated data based on
  #the null probability models are further away from the mean simulated
  #interaction term than the interaction term from the actual
  #experimental results?
  else if(alternative == "two.sided") {
    centered_sim_interaction_terms <-
      scale(sim_betas[, c("Voicing1:Stopping1")],
            scale = FALSE) %>%
      as.vector()
    centered_actual_interaction_term <- actual_interaction -
      mean(sim_betas[, c("Voicing1:Stopping1")])
    prop <- sum(abs(centered_sim_interaction_terms) >
                  abs(centered_actual_interaction_term)) /
      NROW(sim_betas)
  } else {
    stop("alternative argument must be 'one.sided' or 'two.sided'")
  }
  prop
}


exp2_sim_results <-
  data.frame(
    ProbabilityModel =
      c(rep("All generalizations evaluated together", 2),
        rep("One generalization evaluated at a time", 2)),
    Simulation = rep(c("First", "Second"), 2),
    PropSimulatedInteractionsFurtherAway =
```

```
      c(get_prop_further(exp2_evaluated_together_simulated_betas_1,
                         exp2_interaction_term),
        get_prop_further(exp2_evaluated_together_simulated_betas_2,
                         exp2_interaction_term),
        get_prop_further(exp2_evaluated_one_at_a_time_simulated_betas_1,
                         exp2_interaction_term),
        get_prop_further(exp2_evaluated_one_at_a_time_simulated_betas_2,
                         exp2_interaction_term)))


exp3_sim_results <-
  data.frame(
    ProbabilityModel =
      c(rep("All generalizations evaluated together", 2),
        rep("One generalization evaluated at a time", 2)),
    Simulation = rep(c("First", "Second"), 2),
    PropSimulatedInteractionsFurtherAway =
      c(get_prop_further(exp3_evaluated_together_simulated_betas_1,
                         exp3_interaction_term),
        get_prop_further(exp3_evaluated_together_simulated_betas_2,
                         exp3_interaction_term),
        get_prop_further(exp3_evaluated_one_at_a_time_simulated_betas_1,
                         exp3_interaction_term),
        get_prop_further(exp3_evaluated_one_at_a_time_simulated_betas_2,
                         exp3_interaction_term)))


#Table 5 in paper
print(exp2_sim_results)
```

```
#Table 8 in paper

print(exp3_sim_results)
```