Report

Rapid generalization in phonotactic learning: a replication project

Nay San*

June 2020

1 Introduction

Natural languages place various restrictions on how sounds may combine to create words, and phonotactics is the set of these restrictions. For example, in English phonotactics, while the velar nasal sound [ŋ] is permissble word-finally, e.g. *king* [kɪŋ], it is not permissble word-initially, e.g. *[ŋɪk] (even though other nasals are, e.g. *nick* [nɪk]). By contrast, in many other languages word-initial [ŋ] is entirely permissble: e.g. [ŋo] meaning *to cry* in Burmese. Language learners are virtually never given explicit instruction about these restrictions, yet they come to have intuitions about how words in the target language should sound. In essence, language learners automatically form generalisations about the phonotactics of a language based on their exposure to words in the language. How varying amounts of exposure to an artificial language effects phonotactic generalisations about the language was investigated by Linzen and Gallagher (2017) in a series of four experiments. In this report, I present results from a replication experiment based on one of these experiments.

2 Background: Linzen & Gallagher (2017)

Sensitivity to phonotactically legal and illegal sequences of sounds has been shown across a variety of experiments, including artificial language learning experiments. In these experiments, participants are exposed to a set of words from a miniature artificial language in an exposure (or training) stage and in a subsequent testing stage complete a series of judgement tasks to establish what knowledge about the language they have inferred from the exposure stage. Previous studies investigating phonotactic learning, however, typically provide participants with a large amount of exposure to the artificial language. Thus, little is known about the early stages and the time course of phonotactic learning, prompting the questions: How much exposure to the language is needed before learners form generalisations? Are multiple instances of a certain phonological feature required to prompt generalisations? How does the likelihood of generalisation change with more instances?

To gather empirical data towards addressing these questions, Linzen and Gallagher (2017) conducted 4 experiments investigating phonotactic learning. In brief, each subsequent experiment altered the basis on which generalisations may be formed: in Experiment 1, the participants were presented with categorical data (all consonants of interest in the exposure stage shared the relevant phonological feature). In Experiments 2a and 2b, they were probabilistic: a fraction of the data were phonotactically legal. Finally, whether or how participants generalised from a single instance of a sound was investigated in Experiment 3. Given the increased complexity in the composition of artificial languages in each subsequent experiment, more participants were needed in the later experiments (e.g. 450 in Experiment 3 vs. 288 in Experiment 1). Given the constraints of a class project, I chose to replicate Experiment 1, and thus only focus on details of Experiment 1 in the remainder of this section.

*Many thanks to Judith Degen for providing the opportunity and resources to make this class project possible, as well as Tal Linzen and Gillian Gallagher for having made their experiment materials readily available on GitHub along with helpful correspondance during the construction of my replication experiment. Any errors introduced are my own.

2.1 *Materials*

Table 1 provides an overview of the stimuli used by Linzen and Gallagher (2017). To create the artificial languages for the experiment, 540 nonce-words of the form $C_1V_1C_2V_2(C_3)$ were recorded by a native English speaker. The word-initial consonants (the segments of interest) were drawn from a set of six voiced obstruents [b d g v z ð] and six voiceless obstruents [p t k f s θ], whose phonological features are provided in Table (1b).

Table (1c) provides an overview of the counter-balancing design, in which each row of 3 consonants provide the basis of an artificial language, which consists only of voiced consonants (Lists 1-6) or voiceless consonants (Lists 7-12) — n.b. the non-conforming consonant in the third column has the opposite voicing of those in the previous 2 columns (as indiciated by the blue vs. red text colour). Note that while the conforming consonants share voicing, the non-conforming consonant shares the same place of articulation as the unattested but conforming one, e.g. [b] and [p] in List 1 are both labial consonants (as indicated by the underlining).

Table 1: Stimulus design from Linzen and Gallagher (2017). Table (a) describes the composition of the 540 stimulus items. Table (b) lists the phonological features of the consonants of interest, C_1. Table (c) reports the counter-balancing design.

(a)

C_1	V_1	C_2	V_2	(C_3)
b	a	l	a	-
ð	e	m	i	-
d	i	n	u	-
f	o	l	i	-
t	u	n	a	l

where

$C_1 \in \{$ p, t, k, f, θ, s, b, d, g, v, ð, z $\}$
$V_1 \in \{$ a, e, i, o, u $\}$
$C_2 \in \{$ l, m, n $\}$
$V_2 \in \{$ a, i, u $\}$
$C_3 \in \{$ l, m, n $\}$, if $C_1V_1C_2V_2$ is an English word (e.g. *tuna*)

Total stimulus items = $540 = 12 \times 5 \times 3 \times 3$
$= |C_1| \times |V_1| \times |C_2| \times |V_2|$

(b)

Voicing	Manner	Place of articulation: Labial	Coronal	Dorsal
Voiceless	Stop	p	t	*k*
Voiceless	Fricative	f	θ s	
Voiced	Stop	b	d	*g*
Voiced	Fricative	v	ð z	

(c)

List ID	Attested	Unattested: Conforming	Unattested: Non-conforming
1	*g*	b	p
2	z	d	t
3	ð	*g*	*k*
4	d	v	f
5	v	z	s
6	b	ð	θ
7	*k*	p	b
8	s	t	d
9	θ	*k*	*g*
10	t	f	v
11	f	s	z
12	p	θ	ð

2.2 *Procedure*

At the start of each experiment, participants were assigned to a specific list (e.g. List 11) and a specific number of exposures (e.g. 1-set). With these two parameters, stimuli that were compatible with the list were drawn randomly from the pool of 540 items at run time. For example, given that List 11 is a voiceless language where [s] is to be the held out consonant, the remaining 5 voiceless obstruents are [p t k f θ]. Thus, in a 1-set condition, 5 words beginning with these consonants (e.g. [pinu], [tanu], [kelo], [fula], [θomi]) would be randomly drawn for the exposure stage of the experiment. For the 2-, 4-, and 8-set conditions, multiple sets were drawn such that the frequency of occurence of each word-initial consonant remained the same at 20% (e.g. in the 2-set condition with a total of 10 words, there would be exactly two [p/t/k/f/θ]-initial words). Thus the exposure stage may vary from 5 trials (1-set) to 40 trials (8-set).

Regardless of the length of exposure, the testing stage consisted of 6 trials for all participants. For the testing stage, two additional words that begin with the attested consonant were drawn, e.g. [falu] and [femi] for [f] (i.e. not [fula] heard in the exposure stage), along with two that begin with the held-out consonant, e.g. [soma] and [sunu] for [s], and the non-conforming consonant, e.g. [zila] and [zoma] for [z]. In each trial in the testing stage, participants were asked "Does this sound like it could be a word of the language you were listening to?". The presentation order of the stimuli were randomised within both the exposure and testing stages of the experiment.

2.3 *Participants*

For each of the exposure conditions (1-, 2-, 4-, or 8-set), 72 participants were recruited on Amazon Mechanical Turk (MTurk) — six participants for each each of the 12 lists in Table 1c —, yielding a total of 288 participants ($288 = 4 \times 6 \times 12$). Three participants were rejected as their reported native language was not English. Thus Linzen and Gallagher (2017) reported results based on responses from 285 participants.

2.4 *Results*

Figure 1 displays the mean endorsement rates for each exposure condition (1-, 2-, 4-, 8-sets) and stimulus type (conforming-attested, conforming-unattested, or nonconforming-unattested) reported by Linzen and Gallagher (2017). The two key results are that 1) there was no main effect of exposure on endorsement rate and that 2), within each exposure group, the nonconforming-unattested stimuli (blue squares in Figure 1) were endorsed significantly less than conforming-attested stimuli (red circles in Figure 1). Additional results from Linzen and Gallagher (2017) will be discussed below in comparison the results of the replication experiment.

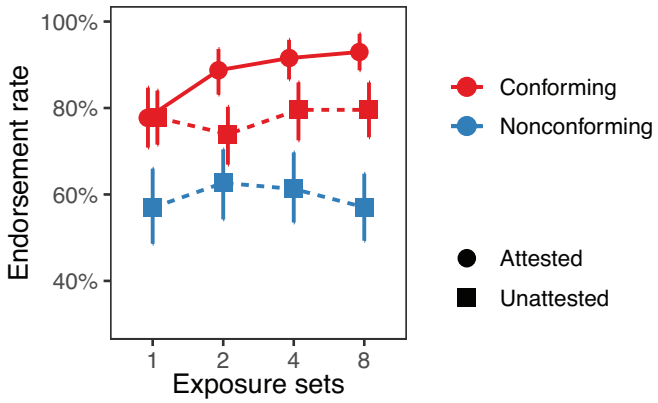

Figure 1: Mean endorsement rates from Experiment 1 of Linzen and Gallagher (2017). Error bars represent bootstrapped 95% confidence intervals.

3 Replication

3.1 Materials, procedure, & participants

All experimental materials used by Linzen and Gallagher (2017) were readily available on a GitHub repository.[1] Thus, the exact same audio stimuli were retrieved for use in the replication experiment. However, as the web experiment framework ExperiGen used in the original experiment was no longer easily deployable, the experiment was adapted into a static site for use with the Stanford ALPS Lab experiment template.[2] While the presentation framework was different, substantive elements of the original experiment were retained (e.g. original JavaScript code for stimulus sampling, original question at test trials: "Does this sound like it could be a word of the language you were listening to?"). All materials used in the replication experiment are also readily available in a GitHub repository,[3] and the experiment was also pre-registered.[4]

Given the resource and time constraints of a class project, and to recuit additional participants per exposure condition, only the 1-set and 4-set exposure conditions were included in the replication experiment (though 109 participants were recruited for each condition, compared to the original 72 per condition). Participants in the 1-set condition were paid $0.50 and those in the 4-set condition were paid $0.75 (calculated from a $15/hour minimum wage and the expected duration of the experiment).

218 participants were recruited from MTurk (129 male, 88 female, 1 other; median age: 35, age range: 18-18, 1 unreported). 5 participants were removed as their reported native language was not English. Thus, the analysis was conducted on data based on 213 participants.

3.2 Analysis

A Bayesian mixed-effects logistic regression was fitted to the data using the `brm` function from the `brms` R package (Bürkner, 2017). Endorsement (0 = does not belong to the language or 1 = belongs to the language) was modelled was a function of number of exposures (1- or 4-set of words; reference level 1), and onset type (conforming-attested, conforming-unattested, or nonconforming-unattested; reference level conforming-attested), and their two-way interaction. The model included the maximal random-effect structure justified by the design: a by-subject intercept and slope for the effect of onset type, and by-onset intercepts and slopes for both onset type and number of exposures. Default priors from the `brms` package were used.[5]

Four sampling chains ran for 4000 iterations with a warm-up period of 2000 iterations for each model, yielding 8000 samples for each parameter tuple. For all relevant cell means and differences between them, the expected values under the posterior distribution and their 95% credible intervals (CIs) are reported. For differences between cells, the posterior probability that a difference δ is not equal to zero is reported. Non-linear hypothesis testing was conducted using the `hypothesis` function from the `brms` package.

4 Results

The mean proportions of test words judged as acceptable by each exposure group for each of the stimulus types are shown in Figure 2, with the left-panel for mean endorsement rates from Linzen and Gallagher (2017) for reference, and the right-panel for results from this replication experiment.

Results from the Bayesian mixed-effects logistic regression indicated that subjects in the 4-exposure condition were not more likely to endorse test words than those in the 1-exposure condition ($\beta = 0.33$, CI = [-0.24, 0.93], $P(\delta > 0) = .867$). In the 1-exposure condition, subjects were less likely to endorse

[1] https://github.com/TalLinzen/rapid_phonotactic_generalization

[2] https://fauxneticien.github.io/phon-learning/experiment/?l1e4, with the parameters `l1` for List 1 and `e4` for 4 exposure sets.

[3] https://github.com/fauxneticien/phon-learning

[4] https://osf.io/jkm2a

[5] I opted to use the default priors since I wasn't sure whether or how to specify explicit priors based on results from Linzen and Gallagher (2017).

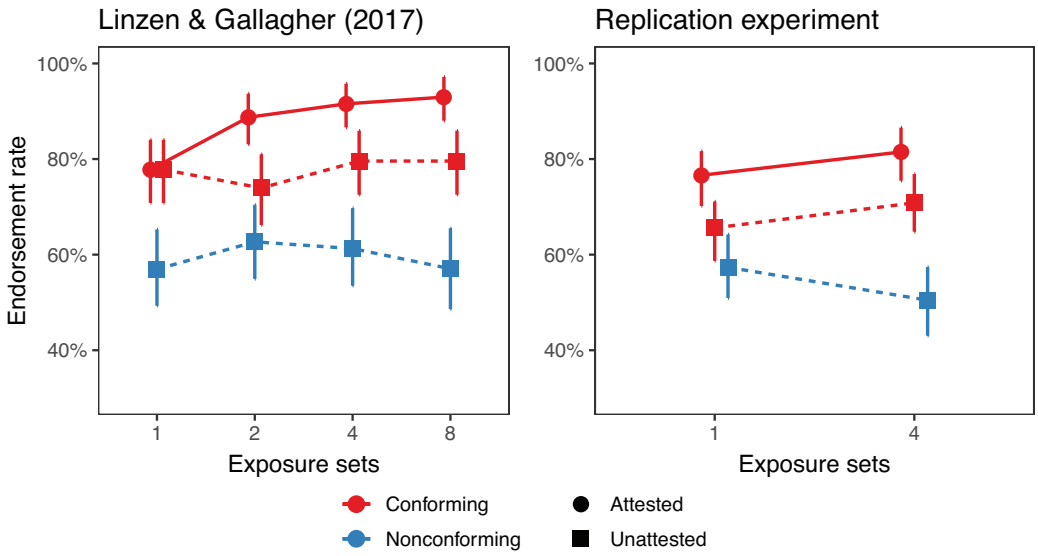

Figure 2: Mean endorsement rates from Experiment 1 of Linzen and Gallagher (2017) (left panel) and current replication experiment (right panel). Error bars represent bootstrapped 95% confidence intervals.

nonconforming and unattested test words than conforming and attested ones ($\beta = -0.95$, CI = [-1.48, -0.39], $P(\delta < 0) = .999$), but did not endorse conforming and unattested test words differently from conforming and attested ones ($\beta = -0.44$, CI = [-1.01, 0.16], $P(\delta < 0) = .930$). Likewise, in the 4-exposure condition, subjects were less likely to endorse nonconforming and unattested test words than conforming and attested ones ($\beta = -1.09$, CI = [-2.11, -0.09], $P(\delta < 0) = .970$), but did not endorse conforming and unattested test words differently from conforming and attested ones ($\beta = -0.37$, CI = [-1.38, 0.62], $P(\delta < 0) = .730$). Finally, while in the 4-exposure condition, subjects were less likely to endorse nonconforming and unattested test words than conforming and unattested ones ($\beta = -0.71$, CI = [-1.4, -0.05], $P(\delta < 0) = .960$), they did not endorse nonconforming and unattested test words differently to conforming and unattested ones in the 1-exposure condition ($\beta = -0.51$, CI = [-1.04, 0.02], $P(\delta < 0) = .940$).

5 Discussion

The original experiment by Linzen and Gallagher (2017) and this replication investigated how participants who listened to varying amounts of stimuli from an artificial language in an exposure stage then judged novel stimuli in a test stage that were identical, similar, or dissimilar to those previously heard. The two key results from Linzen and Gallagher (2017) were that amount of exposure had no significant effect on mean endorsement rate and that, within all exposure conditions, nonconforming and unattested stimuli were endorsed significantly less than conforming and attested stimuli. Both these results were replicated.

Linzen and Gallagher (2017) also found an interaction effect between amount of exposure and stimulus type. While the presence of the interaction between these variables was replicated, the exact interactions were only partially replicated. In both the original experiment and this replication, participants did not endorse conforming and unattested stimuli differently from conforming and attested ones, in neither the 1-set nor 4-set condition. Similarly, in both experiments, participants endorsed nonconforming and unattested stimuli less frequently than conforming and unattested ones in the 4-set condition. While Linzen and Gallagher (2017) found this same contrast in endorsement rates in the 1-set condition, participants in the 1-set condition of the replication experiment did not differ in their endorsement rates of nonconforming and unattested stimuli and conforming and unattested ones.

The partial replication with regards to the interaction between amount of exposure and stimulus type may be related to the highly similar nature of the stimuli, as discussed by Linzen and Gallagher (2017, p. 10). As noted by the authors, all exposure words had the same voicing and were of the CVCV shape with stress on the first syllable. The apparent influence of these similarities seem to be reflected in both the original and replication experiments, where even the nonconforming and unattested stimuli were endorsed relatively highly (around 60%). Linzen and Gallagher (2017, p. 10) noted that "test words that differed from the exposure words in more dimensions, such as *ulpiuzi* or *eh* would have been endorsed at a lower rate". While this experiment could not have added such stimuli, one future direction would be to further investigate additional differences in stimuli and their consequence to endorsement rates in the same limited exposure setting (recall participants in the 1-set listening to only one set of 5 words).

The results of the replication are compatible with the general argumentation of Linzen and Gallagher (2017). That learners consistently endorse dissimilar sounds relatively less frequently than identical sounds after as few as 5 words is incompatible with models of phonotactic learning that posit a minimal exposure requirement. That learners don't appear to differentiate identical sounds from similar ones is incompatible with models that do not posit a minimal exposure requirement but do posit that learners would readily differentiate between such sounds. In this way, this replication experiment contributes additional empirical data to be reconciled by more adequate models which make explicit predictions about the time course of phonotactic learning and its relation to amount of exposure.

References

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Linzen, T. & Gallagher, G. (2017). Rapid generalization in phonotactic learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1). doi:10.5334/labphon.44