Cumulativity by default in phonotactic learning*

Canaan Breiss*

University of California, Los Angeles

Abstract

An ongoing debate in phonology concerns whether the grammar is better characterized by frameworks which use strictly-ranked constraints (such as Optimality Theory, "OT") or weighted constraints (Harmonic Grammar, "HG"). This paper uses a series of Artificial Grammar Learning experiments focused on static phonotactics to probe an empirical domain where OT and HG make different empirical predictions: cumulative constraint interactions, also known as "gang effects". OT does not allow gang effects by default, while HG permits ganging automatically. I show that learners exhibit spontaneously emerging ganging behavior in a poverty-of-the-stimulus environment, providing experimental data supporting weighted-constraint theories of phonological grammar.

Keywords: phonotactics, cumulative constraint interaction, gang effects, poverty of the stimulus, artificial grammar, acquisition

1. Introduction

A longstanding research question in cognitive science concerns whether people make decisions by weighing all available evidence, or via a more in-

*Many thanks to Bruce Hayes and Megha Sundara for their guidance and advice on all aspects of this research. Thanks also to Adam Chong, Adam Albright, Sara Finley, James White, Colin Wilson, and members the audience at the LSA Annual Meeting 2019, SCAMP 2019, and the UCLA Phonology seminar for much helpful discussion. Thanks also to Beth Sturman and Eleanor Glewwe for recording stimuli; to Henry Tehrani for technical assistance; and to Lakenya Riley, Grace Hon, Amanda Singleton, Shreya Donepudi, Azadeh Safakish, and Ruby Dennis for helping to run subjects. Full responsibility for all remaining errors rests with the author. This work was supported by NSF Graduate Research Fellowship DGE-1650604 to Canaan Breiss.

*Corresponding author; cbreiss{@ucla.edu |.com}

formationally constrained criterion which attends to only the most relevant information. Evidence supports both sides of the question; see Gigerenzer and Gaissmaier [1] for an overview.

Within generative linguistics, a version of this debate surrounds decision heuristics in phonology. Two prevalent frameworks make divergent predictions about how speakers use information when determining the outcome of a phonological process. Optimality Theory (henceforth "OT"; Prince and Smolensky [2], *et seq.*) holds that speakers employ an informationally-frugal, heuristic approach: constraints on well-formed phonological structures are strictly ranked, and the choice between any two possible outcomes is determined solely by the highest-ranking constraint that distinguishes the two. The related Harmonic Grammar framework (henceforth "HG"; Legendre et al. [3], *et seq.*) holds that speakers take an informationally-holistic approach, weighing all relevant cues and choosing the globally optimal outcome. This paper presents evidence from a series of experiments which investigate whether a ranked- or weighted-constraint decision heuristic better characterizes the behavior of human learners in the face of ambiguous data. The present study contrasts with previous work in two ways: it uses an artificial grammar learning paradigm to disentangle the effects of lexical statistics from that of the grammar in phonological judgments, and examines cumulative constraint interaction in the learning of phonotactics rather than alternations, a domain which has received less attention but constitutes a strikingly straight-forward test case for phonological theories.

2. Decision metrics in phonology

The difference between the OT and HG decision heuristics can be observed in the schematic tableaux in 1 and 2. In 1, candidate outcome B wins out at the expense of candidate outcome A, indicated by the pointing arrow. This is because candidate A violates the higher-ranked Constraint 1, indicated by the asterisk in the corresponding cell, while candidate B does not. Because Constraint 1 is ranked above Constraint 2 — indicated by the solid line dividing the two columns — candidate A's single violation of Constraint 1 is more important than candidate B's two violations of Constraint 2. This removes candidate A from contention (indicated by the exclamation point following the asterisk), and so candidate B is deemed optimal.

In the Harmonic Grammar tableau in 2, on the other hand, the optimal outcome is the one which has the lowest penalty score, or harmony (H),

	CONSTRAINT 1	CONSTRAINT 2
Candidate outcome A	*!	
→ Candidate outcome B		**

Figure 1: Schematic example of an OT tableau

considering all constraints and violations. Each candidate's harmony is equal to the number of times it violates each constraint, multiplied by the weight of the constraint violated. Under this approach, the same violations as in 1 result in candidate A being deemed optimal because it has a lower harmony score than candidate B. This is because the two violations of Constraint 2, although tolerated individually, together outweigh the penalty associated with the single violation of Constraint 1.

		CONSTRAINT 1	CONSTRAINT 2	
	weights	3	2	$\mathscr{H}$
→	A	*		3
	B		**	4

Figure 2: Schematic example of a HG tableau

The HG and OT decision heuristics make divergent predictions about the same schema because candidates' constraint violations are cumulative in HG but not in OT.

There are two logically-possible types of constraint cumulativity (cf. Jäger and Rosenbach [4]): *counting cumulativity* and *ganging cumulativity*. Counting cumulativity, illustrated above, occurs when one violation of a lower-weighted constraint leads to a lower penalty than one violation of another, higher-weighted constraint, but two violations of the first constraint are together more penalizing than the single violation of the second. Formally, counting cumulativity obtains when the following inequality holds: $2 \times w(\text{Constraint A}) > w(\text{Constraint B}) > w(\text{Constraint A})$, where $w()$ is a function that returns the weight of a constraint. In contrast, ganging cumulativity characterizes weightings where independent violations of low-weighted constraints are less penalizing than a single violation of a higher-weighted constraint, but when they occur together, these lower-weighted violations

"gang up" together to produce a more severe penalty. This weighting can be described as w(Constraint A) + w(Constraint C) > w(Constraint B) > w(Constraint A), w(Constraint C).

2.1. Phonotactics as a testing ground for theories of acquisition

Let us now turn to the phonological processes that these decision methods seek to model. Although much of the debate between OT and HG in phonology has centered on the long-studied relationship between phonological input forms and possible outputs (Goldwater and Johnson [5], Coetzee and Pater [6], Pater [7], Zuraw and Hayes [8], *i. a.*), the same decision rules are also employed when speakers face another relevant analytic task: determining what sorts of sounds and sound-combinations are characteristic of well-formed words their native language — the language's *phonotactics*.

A large body of work has demonstrated that speakers possess sensitive, gradient well-formedness judgments about a range of attested and unattested phonological structures, implying a rich body of implicit phonological knowledge (see, e.g., Scholes [9]; for recent overviews of the empirical and theoretical landscape see Daland et al. [10], Jarosz [11]). This knowledge can be invoked with a simple test: suppose a speaker hears the word *mlep* and is asked if it could be a word of their language (which, for the sake of example, let us suppose is English). In the case of *mlep* it seems clear that their answer will be informed primarily by the highly unusual cluster *ml* with which the word begins. Yet many problems of this form in natural language acquisition will be not so clear-cut: novel words the infant encounters may contain more than one aberrant region, or the same aberrancy in more than one location. These are cases where OT and HG make different predictions about speakers' well-formedness judgments because of possible cumulative constraint interactions. Therefore, differences between HG and OT decision heuristics should be observable in the end-state of phonotactic learning, and thus amenable to investigation in the lab.

2.2. Cumulativity in natural language

This paper is not the first to examine how multiple coincident marked structures affect the grammar. The additive combination of factors influencing the likelihood of *-t/-d* deletion in corpus data was noted by Guy and Boberg [12] as early as 1994, and the difficulty the phenomenon presented for OT was pointed out not long after (Guy [13]). More recently, experimental

evidence has emerged to support the idea that multiple simultaneous phonotactic violations combine to create an additive effect. Using a speech-error elicitation task, Rose and King [14] examined the effect of simultaneously violating several consonant co-occurrence restrictions in two Ethiopian Semitic languages, Chaha and Amharic. They found that participants produced more errors when stimuli violated both constraints than when stimuli violated each constraint independently. Pater [7] analyzes data from Japanese loan words (originally from Nishimura [15]) to argue that a static phonotactic restriction known as "Lyman's Law" which prohibits more than one voiced obstruent within a word can be construed as a case of constraint cumulativity. Pater finds that while speakers tolerate voiced obstruents and geminate consonants unrepaired when adapting loan words, they preferentially repair words which contain voiced geminates by devoicing, thus enforcing the upper limit on voiced obstruents. Kawahara [16, 17, 18] follows this formal analysis up with a series of acceptability judgment studies on linguistically-naïve participants, which show robust support for Pater's conclusions about the graded nature of this effect, and also finds that Lyman's Law violations can block a process of inter-morphemic obstruent voicing known as *rendaku* (Kawahara [19]) in a further case of apparent cumulativity.

Several studies have also investigated the interaction of multiple simultaneous violations on the assessed well-formedness of nonwords. Pizzo [20] finds that novel pseudo-English words like *tlag* and *plavb*, which violate onset- and coda-phonotactic restrictions of English respectively, are rated as more well-formed than words like *tlavb*, which contain both types of violations (and less well-formed than *plag*, which violates neither); work by Albright [21] contains similar findings. Taken at face value, these studies would seem to support the predictions of HG — multiple simultaneously-violated constraints together have an effect on speakers' judgments which is greater than that of each constraint alone.

2.3. Phonotactic acceptability and the lexicon

While suggestive of cumulative behavior, however, these findings have an alternative explanation. This is because in each of these cases, participant-assessed well-formedness is highly correlated with lexical frequency. Even setting aside models which explicitly estimate phonological well-formedness in part on the basis of phonological neighborhood density (ex., the Generalized Neighborhood Model, Bailey and Hahn [22]), the prominent role of lexical statistics in influencing well-formedness judgments is well established

in generative phonology (Frisch et al. [23], Shademan [24], among many others). Pioneering work by Coleman and Pierrehumbert [25] highlighted the connections between the lexicon and phonotactic well-formedness as part of building a model for predicting well-formedness judgments for nonwords. Albright [26], Fukazawa et al. [27], Kawahara and Sano [28] also find evidence for a complex interaction of lexical statistics and phonological acceptability: under-attestation of words in the lexicon which contain *two* marginal structures results in a dramatic decrease in the acceptability of novel structures of this type relative to those containing only one of the structures.

Further, there is evidence that the relationship between lexicon and phonology is diachronically bidirectional: Martin [29, 30] finds that under the assumption people prefer to reuse novel coinages which are phonotactically well-formed, the lexicon eventually comes to underrepresent phonotactically ill-formed words, setting the stage for a possible feedback loop between synchronic phonotactic judgments which are sensitive to lexical statistics, and lexical statistics which are shaped by a synchronic preference for phonotactic well-formedness. Thus in natural languages the question of directionality — whether words are judged to be ill-formed because they are improbable on the basis of lexical statistics, or whether skewed lexical statistics are the product of the phonological grammar — cannot be satisfactorily resolved.

To disentangle the effects of phonotactic acceptability from that of lexical frequency, I used an artificial grammar learning (AGL) paradigm to create a "sandbox environment" where lexical statistics — in this case, the number and distribution of phonemes in the training data to which participants are exposed — can be carefully controlled. AGLs have been used to ask questions about language acquisition and learnability in a variety of domains (Moreton and Pater [31, 32], Culbertson [33], Finley [34]). In the domain of phonotactics, the AGL paradigm has typically been used to compare typologically attested and unattested phonological patterns to see if differences in attestedness in typology could be due to differing levels of formal learnability of the patterns (Moreton [35], Glewwe [36]). They have been used considerably less often to investigate the kinds of hypotheses that are entertained about the language during learning (although see Linzen and Gallagher [37]). Here I use the paradigm to explore whether knowledge about well-formedness of different types are evaluated cumulatively when generalizing to novel words. I will take participants' inferences about such (non)cumulativity — made in the absence of disambiguating evidence — to be revealing of the "default" procedure with respect to phonotactic learning. This finding, then, can be

used to compare OT and HG decision heuristics as suitable models of the phonological grammar.

2.4. Experimental design choices

All experiments in this paper paired varieties of consonant and vowel harmony. These phenomena have traditionally constituted a core object of generative phonological analysis (see Hansson [38] and Walker [39] for overviews of empirical data on many consonant and vowel harmony patterns respectively), and both have been successfully learned in other AGL experiments (ex., Finley [40], Lai [41]). Because the current experiments focused on simultaneous acquisition of two separate phonotactic patterns, it was crucial that the aspects of the artificial language governed by each of the phonotactics not overlap: consonant harmony regulated all and only a word's consonants, and vowel harmony regulated the vowels, allowing a word to conform to or violate each phonotactic independently. In Experiments 1, 3a-b, and 4 I used a nasal consonant harmony (hereafter *nasal harmony*): a word's consonants agree in nasality, being either drawn from the set of nasal stops {/m, n/} or voiceless stops {/p, t/}; for a survey of parallels in natural languages, see Hansson [38, p. 111 *et seq.*]. In Experiment 2 I used sibilant consonant harmony (*sibilant harmony*): consonants in a word agree in anteriority, being drawn from {/s, z/} or {/ʃ, ʒ/} (see Hansson [38, p. 55 *et seq.*] for a typological survey). All experiments presented here used vowel backness harmony (*backness harmony*): all vowels in a word agreed in backness, being {/i, e/} or {/u, o/} (for an overview, seeWalker [39]).

To ensure an accurate assessment of participants' well-formedness judgments, I elicited acceptability judgments from participants using two different tasks. First, participants completed a *lexical decision task* in which they were asked to judge whether a novel would could belong to the language that they learned at the start of the experiment (possible answers *yes* or *no*). They then completed a *ratings task* where they were asked to assign each of those same words a numerical rating (scale from 0 (*very bad*) to 100 (*very good*)) based on how good that word sounded as an example of the language they had learned. I collected both threshold-based categorical data and continuous data to determine whether participants assessments of well-formedness pattern categorically (as would be expected if participants only attended to one cue to phonotactic well-formedness, as in OT), or are distributed over a gradient of well-formedness (as predicted by HG). Solid evidence in favor of either outcome should be the result of converging evidence from both

dependent variables.

3. Experiment 1

In Experiment 1, I tested whether learners inferred a cumulative effect between violations of two different phonotactics — *ganging cumulativity*.

3.1. Methods

3.1.1. Participants

45 undergraduate students at a North American university were recruited to participate in this experiment through the Psychology Subject Pool, and were compensated with course credit. Participants who had not spoken English consistently since birth were excluded ($n = 2$), as were those who did not meet the criterion for learning assessed during the verification phase ($n = 10$, on which more below), leaving 33 participants whose data were included in the final analysis.

3.1.2. Stimuli

In exposure phase, subjects heard 32 initially-stressed ˈCVCV nonwords which conformed to the nasal harmony and backness harmony phonotactics. Individual consonant and vowel identity was balanced in frequency and distribution over word positions. This procedure yielded a language containing words such as *potu*, *meni*, *nuno*, *tepi*, *teti*, *mumo*, etc.

For the verification phase I created two sets of items, each set consisting of 16 pairs of minimally-differing nonwords. One member of each pair was a fully-conforming word from the exposure phase, and the other was created by reversing the featural specification for backness or nasality of one of the consonants or vowels in the fully-conforming word. Thus the pair of words differing only in a single instance of that feature. In each set, 8 pairs differed in a violation of nasal harmony, and 8 differed in a violation of backness harmony, with differences between pair-members balanced for segmental placement and identity. For example, the familiar word *potu* was modified by altering the nasality specification of its second consonant, yielding the pair *po*$\underline{t}$*u* vs. *po*$\underline{n}$*u*.

In piloting, participants showed a strong preference for forms with identical consonants or vowels despite no numerical advantage for these forms in the training data. Therefore, the verification trials were balanced so that pairs whose fully-conforming word had identical consonants (ex. *totu*) differed only

in their violation of backness harmony (ex., *totu* vs. *toti*). On trials whose conforming word contained identical vowels, the two words differed only in a violation of nasal harmony. Crucially, there were no doubly-violating words in the verification phase: the purpose was simply to ensure that subjects had learned each of the two phonotactic constraints independently.

In the test phase, subjects were presented with a set of 48 novel non-words which varied in conformity both phonotactics. 24 conformed to both phonotactics (ex. *pite*), eight violated only the nasal-harmony phonotactic (*mite*), eight violated only the backness-harmony phonotactic (*pito*), and eight violated both the nasal-harmony and backness-harmony phonotactics (*mito*).

All words were recorded by a phonetically-trained female native English speaker using PCQuirer. They were digitized at 44,100 HZ and normalized for amplitude to 70 db.

3.1.3. Design

The experiment consisted of an exposure phase followed by a verification phase, after the successful completion of which participants moved on to two successive generalization tasks in the test phase: the lexical decision task and then the ratings task. The exposure phase consisted of two blocks of 32 pseudo-randomized self-paced trials. During the exposure phase, similar-sounding items were presented together in blocks of eight, with each subject assigned at random to one of four counter-balancing orders of the four blocks. For example, in one counterbalancing group, participants first heard eight words with front vowels and voiceless stops (ex. *peti, tipi, tepe, piti...*) followed by eight words with back vowels and nasal stops (*monu, nunu, mumo, numo...*), followed by eight words with back vowels and voiceless stops (*topu, pupo, topo, putu...*) followed by eight words with front vowels and nasal stops (*nini, meni, nemi, mene...*).

After the exposure phase, participants completed 16 self-paced two-alternative forced choice verification trials. If participants scored above 80% (13 or more correct answers) they moved on to the test phase. Otherwise they received another block of 32 pseudo-randomized trials in the exposure phase, after which they completed a second verification phase. The two sets of 16 verification-phase pairs alternated in successive verification phases to lower the likelihood of participants passing verification via trial and memorization alone. If participants did not meet criteria within three additional exposure blocks they were simply asked to complete the demographic questionnaire,

and did not complete the test phase.

If subjects met criteria on the verification phase, they advanced to the test phase which consisted of the lexical decision task and the ratings task. Both tasks used the same set of novel words.

In the lexical decision task, participants were presented with two repetitions of 48 novel words in a random order and were asked to choose whether they thought each word could belong do the exposure language they had learned. In the ratings task, participants were asked to rate each of the same words on a scale from 0 (*very bad*) to 100 (*very good*) based on how good they sounded as an example of the language they had learned in the exposure phase. At the end of the experiment, demographic information was collected. The full experiment lasted approximately 15-20 minutes, depending on the number of exposure blocks the subject required.

3.1.4. Procedure

The experiment was conducted in a sound-attenuated room using a modified version of the Experigen platform (Becker and Levine [42]). After giving their informed consent to participate in the study, at the start of the experiment participants were told that they would be learning a new language, after which they would be tested on their knowledge. Participants were encouraged to repeat back each word they encountered in the experiment to help them get a better sense of the language: both hearing and speaking the words was intended to make the phonotactic patterns more salient and help participants stay focused on the task. Participants were instructed to base their decisions on what they knew about how the language sounded and what their gut told them was right, and to not over-think their choices.

The experiment had a fully self-paced design. On each trial of the exposure phase participants were instructed to click a button on the screen to hear a word of the language. When they did so, they heard one of the 32 fully-conforming words chosen for the exposure phase, sampled without replacement, and were instructed to repeat the word out loud. The verification phase had a similar structure, except each trial played a pair of words in a random order, and participants were instructed to say both words out loud before making their choice. The test phase had a similar structure, with each task consisting of a series of trial containing one word which participants were instructed to repeat out loud before either making the lexical decision or assigning it a rating.

3.2. Analysis

Data from the test phase was analyzed using mixed effects regression models in R (Team et al. [43]) using the *lme4* package (Bates et al. [44]). All statistical analyses began by fitting a maximally-specified model (following Barr et al. [45]), which contained a random intercept for subject and item, fixed effects specific to the analysis, and random slopes for all fixed effects by subject. In cases of non-convergence, interactions among random slopes were removed first, then the slopes themselves, until the model converged. For the lexical decision task, I modeled the log odds of endorsing an item using logistic regression, and for the ratings task I analyzed the raw numerical data using a linear model. Note that although I regressed on the raw ratings data, for the sake of legibility I plot the z-normalized ratings throughout the paper.

3.3. Results I: Do subjects infer ganging cumulativity between phonotactic violations?

The first question (regardless of dependent variable) is simply whether doubly-violating forms are judged worse than singly-violating forms. For this analysis, the status of the form in question (fully-conforming, violating backness harmony, violating nasal harmony, or doubly-violating) was coded as a four-level fixed effect. Holding each level of the fixed-effects as the reference level in turn, I probed whether the log odds of acceptance (in the lexical decision task) or the numerical rating (in the ratings task) for singly-violating levels differed significantly from the fully-conforming level. Significant differences on this metric are a prerequisite for analyzing the difference between singly- and doubly-violating levels, as they indicate learning of the individual phonotactics in independent contexts. The critical comparison for determining whether learners infer a gang effect is between the singly-violating levels and the doubly-violating level, which was probed in the same manner. A significant difference constitutes evidence for a gang effect between the independent phonotactic violations reflected in assessed well-formedness. In contrast, no significant differences between these levels is evidence that learners did not infer a cumulative effect from multiple simultaneous constraint violations.

3.3.1. Lexical decision task

Figure 3 shows the results of the lexical decision task in Experiment 1. The final logistic regression model included a random intercept for sub-

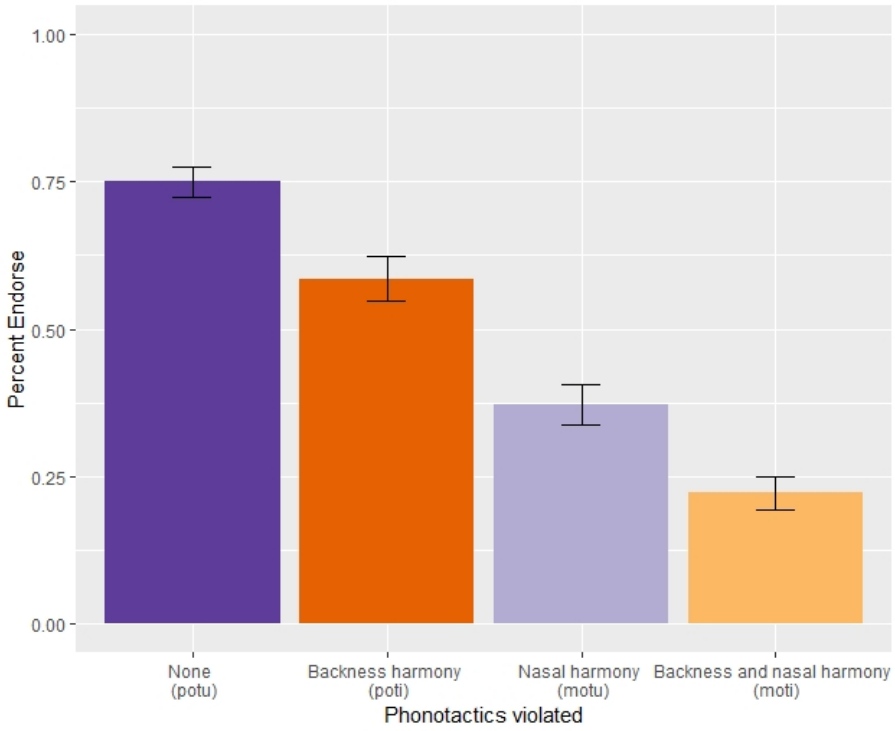

Figure 3: Results for the lexical decision task, Experiment 1. The vertical axis plots mean endorsement rate as a percentage, with standard error bars, and the horizontal axis divides the novel words according to their phonotactic violation profile, together with an illustrative example of that profile type.

ject and word, and a four-level fixed effect for violation profile. The log odds of endorsement differed significantly between the fully-conforming forms and the nasal harmony-violating forms ($\beta = -1.748$, $p < 0.001$), and the backness harmony-violating forms ($\beta = -0.813$, $p < 0.001$). Participants endorsed doubly-violating forms with a significantly lower likelihood than nasal harmony-violating forms ($\beta = -0.793$, $p = 0.002$) and backness harmony-violating forms ($\beta = -1.728$, $p < 0.001$).

3.3.2. Ratings task

Results of the ratings task are presented in Figure 4. The final model included a four-level fixed effect of violation profile, and a random intercept for subject and word. Forms violating the backness harmony phonotactic were rated significantly lower than fully-conforming forms ($\beta = -7.100$, $p =$

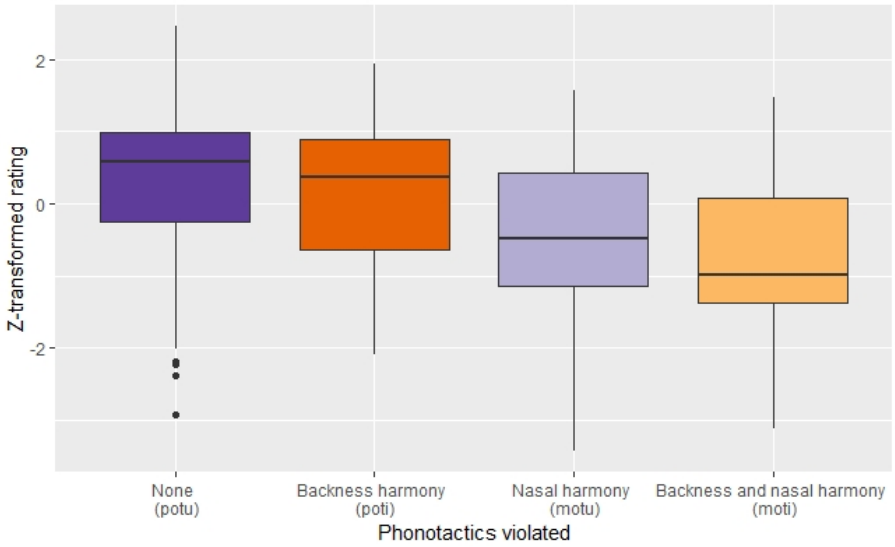

Figure 4: Results for the ratings task, Experiment 1. For readability, I plot z-normalized rating on the vertical axis, and the horizontal axis divides the novel words according to their phonotactic violation profile, together with an illustrative example of that profile type.

0.019) and forms violating the nasal harmony phonotactic were rated significantly lower than fully-conforming forms as well ($\beta = -23.678$, $p < 0.001$). Doubly-violating forms were rated significantly lower than those violating just the backness harmony phonotactic ($\beta = -25.700$, $p < 0.001$) and those which violated just the nasal harmony phonotactic ($\beta = -9.125$, $p = 0.014$).

3.4. Results II: Is the gang effect sub-linear, linear, or super-linear?

The previous analysis finds that participants inferred ganging cumulativity between constraint violations, as predicted by HG, we can further probe the nature of this effect. Is the effect of multiple simultaneous violations linearly additive (doubly-violating words are judged as marked *in proportion* to the sum of their single violations), sub-linear (doubly-violating words are *less marked* than the penalty associated with their isolated violations), or super-linear (doubly-violating words are *more marked* than the sum of their violations)? Unlike the clear predictions of OT and HG about the existence of constraint cumulativity in the general case, there is no clear consensus among HG frameworks about the linearity of such effects. Although

the original formulation of HG (Legendre et al. [3]) predicts only linearly-additive cumulativity, other frameworks that use weighted constraints make different predictions about the specific conditions under which linearity (or lack thereof) should be observed. Since this topic is the subject of ongoing quantitative and theoretical investigation, the linearity of any observed cumulativity will be statistically assessed so it may contribute to this literature, but will not form the basis of any further theoretical claims.

To probe the linearity of the gang effect, a second statistical analysis was carried out on the same data sets using a mixed effects regression model containing two fixed effects — whether a word violated the nasal-harmony phonotactic (*true*, *false*), whether a word violated the backness-harmony phonotactic (*true*, *false*) — and their interaction. Of interest here is the interaction term: a significant interaction with a positive coefficient would indicate that the gang effect inferred was *sub-linear* — two coincident violations were judged less severe than would follow from the combination of their markedness when occurring in isolation. Alternatively, a significant interaction with a negative coefficient would indicate that the gang effect was *super-linear* — two coincident violations were judged more severe than would follow from their independent status in the grammar. Finally, a non-significant interaction would be consistent with a *linear* gang effect: doubly-violating forms are judged ill-formed in proportion to their component violations.

Given that two response variables were collected — binary choice data as well as continuous ratings data — the question arises of how we should interpret any divergences between these two metrics in terms of linearity. Previous work comparing binary and continuous measures of acceptability has found that the two convey generally very similar information (Weskott and Fanselow [46]), and thus we might expect the two measures to yield converging measures of linearity; this issue is taken up again in section 7.2.

3.4.1. Lexical decision task

The final logistic regression model had the two fixed effects and their interaction as described above, and random intercepts for subject and word. Both main effects were significant (nasal harmony violation = *true*: $\beta = -1.747$, $p < 0.001$, backness harmony violation = *true*: $\beta = -0.812$, $p < 0.001$) and the interaction term was not significant ($\beta = 0.020$, $p = 0.951$).

3.4.2. Ratings task

The analysis was performed on the ratings task data using a linear model with two fixed effects and their interaction as described above, and random intercepts for subject and word. The model revealed that both main effects were significant predictors of rating (nasal harmony violation = *true*: $\beta = -23.676$, $p < 0.001$, backness harmony violation = *true*: $\beta = -7.100$, $p = 0.019$) and that the interaction was not ($\beta = -2.025$, $p = 0.663$).

3.5. Local discussion

In summary, Experiment 1 provides evidence that in a poverty-of-the-stimulus environment, learners infer a gang effect between violations of two separate phonotactic constraints in their learning data. In the lexical decision task, doubly-violating forms were endorsed in proportion to the likelihood of endorsement of forms bearing each of their violations independently, and in the ratings task words received a rating proportional to the summed penalty of their independent violations, indicating that the gang effect inferred is linear. This finding is broadly compatible with the predictions of HG models of phonological grammar, and not with those of OT.

4. Experiment 2

To establish the generality of the results of Experiment 1, in Experiment 2 I replicated Experiment 1 with a different consonant harmony phonotactic — *sibilant harmony*.

4.1. Methods

4.1.1. Participants

84 undergraduate students were recruited to participate in this experiment, none of whom had participated in Experiment 1. Participants were excluded for not having spoken English since birth ($n = 15$), and not consistently learning both phonotactic constraints ($n = 35$), leaving 34 participants whose data were included in the study. Recruitment method, compensation, experimental setting and software were the same as for Experiment 1.

4.1.2. Materials

New materials were created for Experiment 2 by systematically altering the stimuli from Experiment 1 in the following way: all cases of /p/ were replaced by /ʃ/, /m/ by /ʒ/, /t/ by /s/, and /n/ by /z/.

4.1.3. Design, procedure, and analysis

Design, procedure, and analysis for Experiment 2 was identical to that of Experiment 1, except that the lexical decision task contained only one presentation of each of the novel words, rather than two.

4.2. Results I: Do subjects infer ganging cumulativity?

4.2.1. Lexical decision task

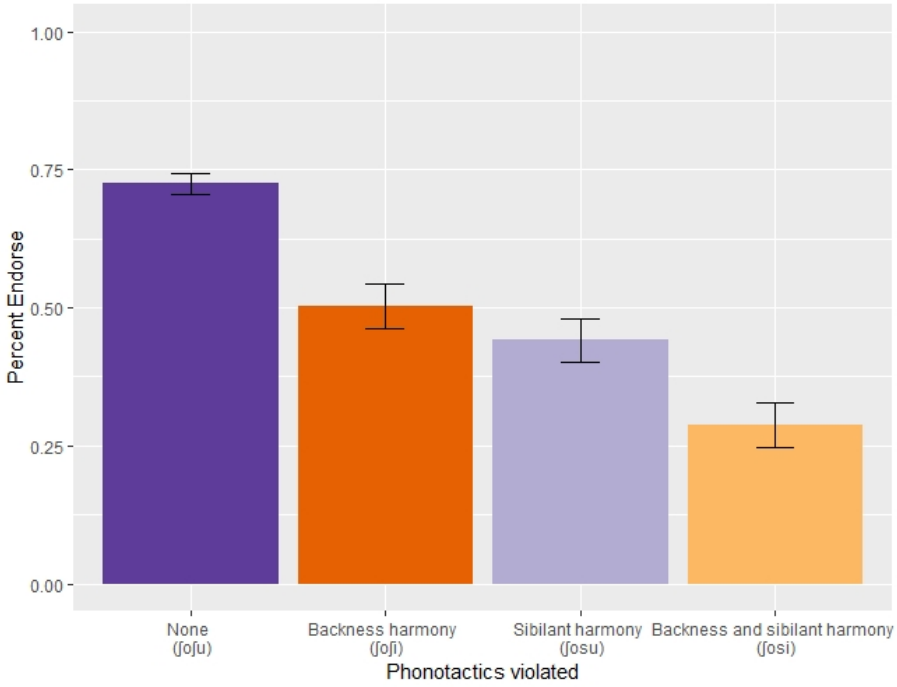

Figure 5: Results for the lexical decision task, Experiment 2.

Figure 5 shows the results of the lexical decision task in Experiment 2. The final logistic regression model contained a random intercept for subject and word and a fixed effect for violation profile. The log odds of endorsement for forms violating only sibilant harmony were significantly lower than that of fully-conforming forms ($\beta = -0.968$, $p < 0.001$), as were the log odds of endorsement for backness harmony-violating forms ($\beta = -1.223$, $p < 0.001$), paralleling the results from the lexical decision task in Experiment 1. Also in line with Experiment 1, participants endorsed words violating both sibilant and backness harmony at lower rates than those which violated only sibilant

harmony (β= −0.940 , $p < 0.001$) or only backness harmony (β= −0.685, $p = 0.002$).

4.2.2. Ratings task

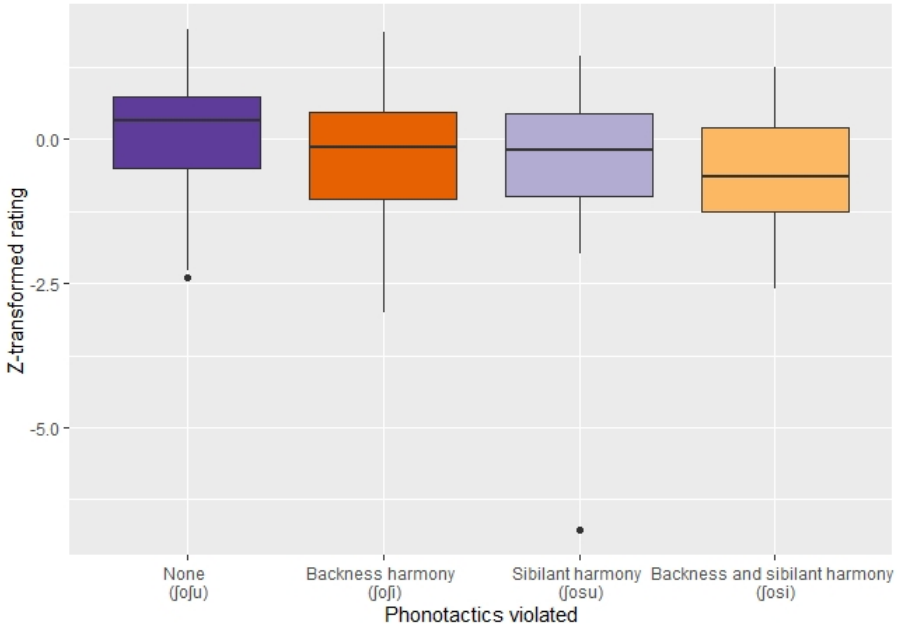

Figure 6: Results for the ratings task, Experiment 2

Results of the ratings task are presented in Figure 6. The final regression modeled raw ratings as a function of violation profile with random intercepts for subject and word. Mirroring the results from the lexical decision task, forms violating only sibilant harmony were rated significantly lower than fully-conforming forms (β= −14.722, $p < 0.001$), as were forms violating only backness harmony (β= −14.429, $p < 0.001$). Doubly-violating forms received lower ratings than those violating only sibilant harmony (β= −15.438, $p < 0.001$) and only backness harmony (β= −15.730, $p < 0.001$), confirming that learners inferred a cumulative "ganging up" effect between constraint violations.

4.3. Results II: Is the gang effect sub-linear, linear, or super-linear?

4.3.1. Lexical decision task

I analyzed the log odds of endorsing a novel word using a logistic regression model with three fixed effects: sibilant harmony-violation (*true*, *false*), backness harmony-violation (*true*, *false*), and their interaction. The model also included random intercepts for word and subject, with by-subject slopes for the independent effects of sibilant harmony-violation and backness harmony-violation. The fixed effects of violating sibilant harmony ($\beta = -1.003$, $p < 0.001$) and backness harmony ($\beta = -1.285$, $p < 0.001$) were significant, and their interaction term was not significant ($\beta = 0.241$, $p = 0.427$).

4.3.2. Ratings task

To analyze the ratings data, I fit a linear model with sibilant harmony-violation, backness harmony-violation, their interaction, and random intercepts for subject and word. Results indicated that while violating each of the phonotactics independently contributed to lower ratings (sibilant harmony-violation = *true*: $\beta = -14.722$, $p < 0.001$; backness harmony-violation = *true*: $\beta = -14.429$, $p < 0.001$), the interaction of the fixed effects was not significant ($\beta = -1.008$, $p = 0.790$), meaning that the penalty for doubly-violating forms was predictable based on the penalty of the individual violations alone.

4.4. Local discussion

The results of Experiment 2 establish the generality of the findings of Experiments 1, confirming that speakers infer ganging cumulativity among several different types of constraint.

5. Experiment 3a-b

Experiments 1 and 2 established that learners consistently infer gang effects between multiple independent constraint violations in phonotactic learning, supporting the predictions of the HG decision heuristic. Crucially, these experiments used an interactive training phase: participants were presented with words in isolation, and instructed to repeat them back aloud. This process stands in stark contrast to the process of natural language acquisition, where "training" is entirely implicit, and phonotactic knowledge is (at least in early infancy) gleaned completely from passive exposure. Therefore, Experiment 3a-b sought to replicate the results of Experiments 1 and 2

using a passive exposure training paradigm designed to more closely mimic first language acquisition.

Experiment 3a

5.1. Methods

5.1.1. Participants

76 undergraduate students were recruited to participate in this experiment, none of whom had participated in Experiments 1 or 2. Participants were excluded for not having spoken English since birth ($n = 10$), and not reliably learning both phonotactics ($n = 0$), leaving 66 participants whose data were included in final analysis. The sample size was increased to compensate for the less controlled nature of the training phase, described below, which left more room for variable strength of learning by individual participants. Recruitment method, compensation, experimental setting and software, and materials were the same as for Experiment 1.

5.1.2. Design

Design for Experiment 3a was the same as for Experiment 1 except as follows. The exposure phase consisted of each of the 32 training words presented in a random order twenty times in a continuous speech stream. Further, since the exposure was designed to be naturalistic, I did not impose an absolute threshold for advancement to the test phase, in contrast to the 80% criterion from the previous experiments. Rather, participants were allowed to advance to the test phase if they did not learn each of the phonotactics to significantly different degrees. This was operationalized by imposing a condition that the difference in number of correct answers between pairs differing only in a nasal harmony violation and those differing only in a backness harmony violation was not allowed to be greater than 3, chosen by using Fisher's exact test (Fisher [47]) to determine the level at which the proportion of correct answers for each phonotactic significantly differed, across the range of possible accuracies. In practice, the passive exposure training lead to performance on the verification phase comparable to that achieved using the more interactive training method (mean accuracy 81.3%), and no subjects were excluded for not having learned both phonotactics to criterion.

Because of the longer exposure phase, the lexical decision task in the test phase consisted of only one randomized presentation of each of the 48 novel words, rather than two as in Experiment 1.

5.1.3. Procedure and analysis

Procedure for Experiment 3a was identical to that of Experiment 1, except that during exposure participants were instructed that they should simply sit and listen to the speech stream, and, if they began to get bored, they should to try to count how many unique words they heard in the speech stream (this task was suggested to encourage participants to attend to the training stimulus). The exposure phase lasted around ten minutes, and the entire experiment took approximately 20-30 minutes, depending on the number of exposure blocks the subject required. Analysis was identical to that of Experiment 1.

5.2. Results I: Do subjects infer ganging cumulativity?

5.2.1. Lexical decision task

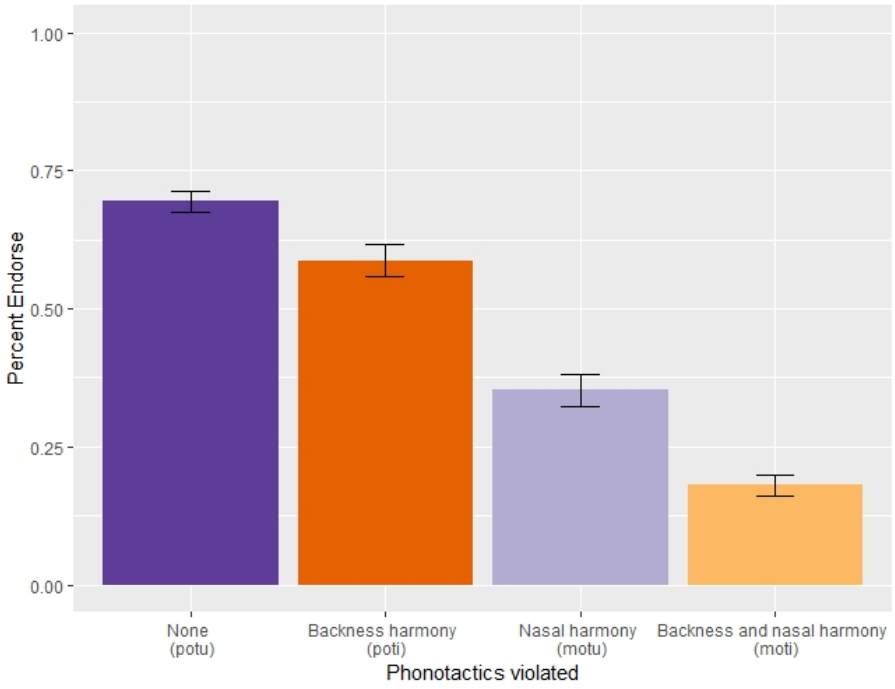

Figure 7: Results for the lexical decision task, Experiment 3a.

Figure 7 shows the results of the lexical decision task in Experiment 3a. The final logistic regression model contained a random intercept for subject and word, and a four-level fixed effect of violation profile. Forms violating

only backness harmony had significantly lower log odds of endorsement compared to fully-conforming forms (β= −0.504, $p = 0.026$), as did those which violated only nasal harmony (β= −1.562, $p < 0.001$). Forms violating both backness harmony and nasal harmony had significantly lower log odds of endorsement than those violating only backness harmony (β= −2.031, $p < 0.001$) and those violating only nasal harmony (β= −0.973, $p < 0.001$).

5.2.2. Ratings task

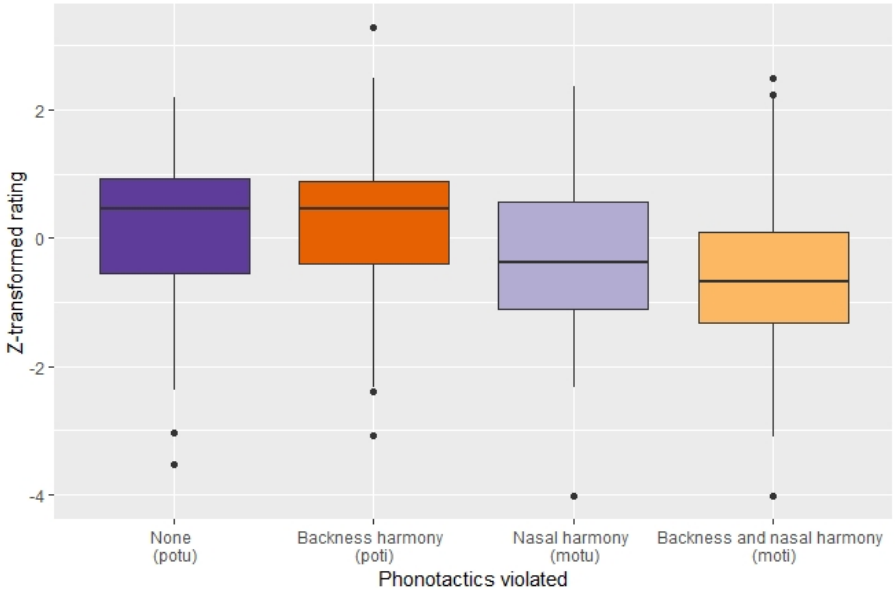

Figure 8: Results for the ratings task, Experiment 3a.

Results of the ratings task are presented in Figure 8. Here, only nasal harmony-violating forms were rated significantly lower than fully-conforming word (β= −17.536, $p < 0.001$); backness harmony-violating forms did not differ significantly in their ratings from fully-conforming words (β= 0.706, $p = 0.850$). Further, doubly-violating words did not differ significantly from nasal harmony-violating words (β= −7.960, $p = 0.086$), although they did differ from backness harmony-violating words (β= −26.202, $p < 0.001$). Since the first analysis did not reveal that learners reliably rated backness-violating forms worse than fully-conforming forms — and thus did not exhibit constraint cumulativity — the test for linearity of gang effects was not carried out on the ratings data.

5.3. Results II: Is the gang effect sub-linear, linear, or super-linear?

5.3.1. Lexical decision task

As before, I analyzed the log odds of endorsing a test item in a logistic regression model with a random intercept for subject and word, and three fixed effects: nasal harmony violation (*true*, *false*), backness harmony-violation (*true*, *false*) and their interaction. I found a significant main effect of nasal harmony violation ($\beta = -1.562$, $p < 0.001$), and a significant main effect of backness harmony violation ($\beta = -0.504$, $p = 026$); the interaction term was not significant ($\beta = -0.469$, $p = 0.196$).

5.4. Local discussion

Experiment 3a provides some evidence for the robustness of the gang effect under more naturalistic passive exposure training: in the lexical decision task, violations of both the nasal harmony and backness harmony phonotactics contributed independently and linearly to likelihood of endorsement. In the ratings task, however, only violations of the nasal harmony phonotactic contributed to lower ratings on average. What could be the cause of this apparent case of learners failing to infer a gang effect between constraint violations? I hypothesize that this effect is due to the change in training paradigm, and concomitant adjustment of verification standards. Passive exposure training may have resulted in more variable learning outcomes, and reduced strength of learning overall, which could have contributed to the quicker decay for the backness harmony phonotactic compared to the nasal harmony phonotactic. Support for this hypothesis comes from the higher accuracy in the verification phase on nasal harmony violating words (mean accuracy 86.7%) compared to backness harmony violating words (mean accuracy 76.1%).

Experiment 3b

To test the post-hoc hypothesis that the lack of a gang effect in the ratings task of Experiment 3a was a task effect, a shortened, ratings-only version of the same experiment was carried out. If participants truly fail to infer a gang effect under more naturalistic training conditions, the results on the ratings task in Experiment 3b should mirror those of Experiment 3a. However, if the results of the ratings task in Experiment 3a were simply due to interference from intervening time and exposure to novel nonconforming

words, the results in Experiment 3b should mirror the results of ratings tasks from Experiments 1 and 2.

5.5. Methods

78 new undergraduate students were recruited to participate in this experiment. Participants were excluded for not having spoken English since birth ($n = 7$), not completing the demographic survey ($n = 1$), and not consistently learning both phonotactics ($n = 0$), leaving 70 participants whose data were included in the study. Recruitment method, compensation, experimental setting and software were the same as for Experiment 1.

Materials, procedure, design, and analysis were identical to those of Experiment 3a, with the exception that participants in this experiment did not complete the lexical decision task during the test phase.

5.6. Results I: Do subjects infer ganging cumulativity?

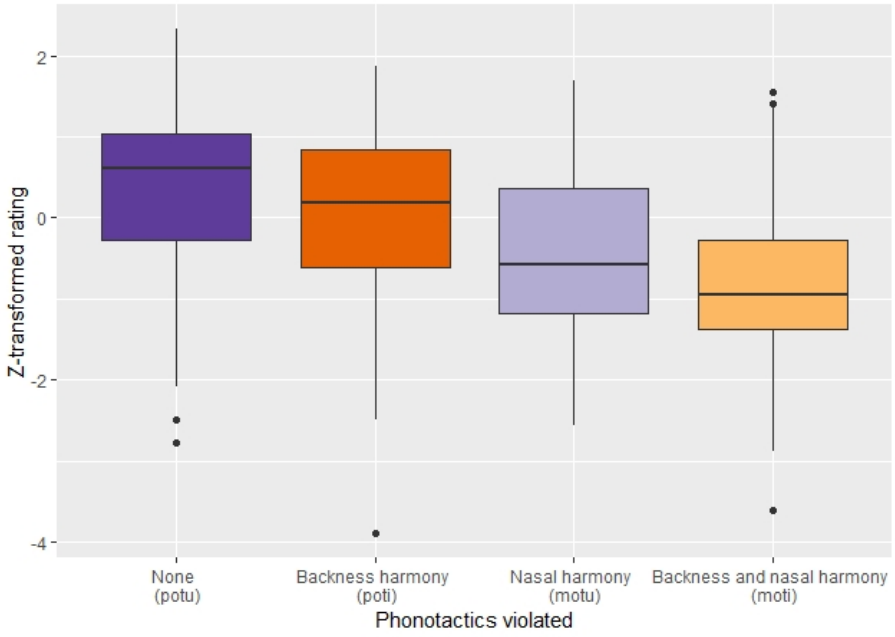

Figure 9: Results for the ratings task, Experiment 3b.

Results of the ratings task are presented in Figure 9. Mirroring results of the lexical decision task from Experiment 3a, forms which violated only

backness harmony received lower ratings than fully-conforming forms (β= −10.136, $p = 0.012$), as did forms which violated only nasal harmony (β= −27.029, $p < 0.001$). Additionally, doubly-violating forms received lower ratings than those which violated backness harmony only (β= −26.725, $p < 0.001$), as well as those which violated nasal harmony only (β= −9.832, $p = 0.044$), confirming that learners inferred ganging cumulativity between constraint violations.

5.7. Results II: Is the gang effect sub-linear, linear, or super-linear?

I fit a linear mixed effects model to the ratings data, with fixed effects of violating backness harmony, violating nasal harmony, and their interaction, along with random intercepts for subject and word. It indicated that violation of the backness harmony phonotactic (β= −27.029, $p < 0.001$) and of the backness harmony phonotactic (β= −10.136, $p = 0.012$) was independently associated with lower ratings, and the interaction between these factors was not significant (β= 0.304, $p = 0.961$). This finding is in line with the lexical decision task from Experiment 3a, as well as Experiment 1 and 2 more broadly.

5.8. Local discussion

Experiment 3b confirms the hypothesis about the apparent lack of a gang effect observed in the ratings task in Experiment 3a: participants likely simply did not recall the language they had learned strongly enough to provide accurate well-formedness judgments in the ratings task.

6. Experiment 4

In Experiments 1, 2, and 3a-b, I examined how single violations of different constraints interact in the grammar — testing for ganging cumulativity. In Experiment 4, I examined the other type of constraint interaction predicted by HG and not by OT, *counting cumulativity*. Because HG takes into account all violations of each constraint, it also predicts that a word violating the same constraint n times will be less well-formed than a near-identical word violating it $n-1$ times. To test this prediction, participants were tested on longer novel words which allowed for each word to host up to two violations of one phonotactic constraint, in addition to violations of the other. This allowed us to test whether counting and ganging cumulativity obtained simultaneously, since we examined number of violations (0, 1, and

2) of each single phonotactic in the context of each level of the other, yielding a fully-crossed design.

6.1. Methods

6.1.1. Participants

71 undergraduate students were recruited to participate in this experiment, none of whom had participated in previous experiments reported here. Participants were excluded for not having spoken English since birth ($n = 12$), not completing the demographic survey ($n = 1$), and not consistently learning both phonotactics ($n = 0$), leaving 58 participants whose data were included in the study. Recruitment method, compensation, experimental setting and software were the same as for Experiment 1.

6.1.2. Materials

Training and verification materials were the same as for Experiment 1. 48 novel test words were created for this experiment, each four syllables long with ˈCVCVˌCVCV syllable structure and a left-aligned trochaic stress pattern (ex., *minemeni*, *putotupo*, *petipite*, etc.). 24 of these words conformed to both nasal harmony and backness harmony phonotactics, and the remaining 24 were divided evenly among the two violation levels (one violation vs. two) of both phonotactics.

6.1.3. Design and procedure

Design for Experiment 4 was identical to that of Experiments 1. Procedure for Experiment 4 was identical to that of Experiment 2, except that participants were instructed before beginning the test phase that they would be tested on longer words, and that even though the words they would be hearing would be longer than the ones they had learned initially, their length did not bear on whether they were likely to belong to the language or not. This point was stressed via an analogy to English, which contains valid words of many lengths.

6.2. Analysis

The analysis of Experiment 4 differed from that of previous experiments because of its design, and because of the additional focus on determining whether learners infer counting cumulativity between constraint violations. I assessed whether learners infer counting cumulativity by holding each level of violation for each phonotactic as the reference level in turn, and noting

whether differences between numbers of violations of the same phonotactic were significantly different.

6.3. Results I: Did learners infer counting cumulativity?

6.3.1. Lexical decision task

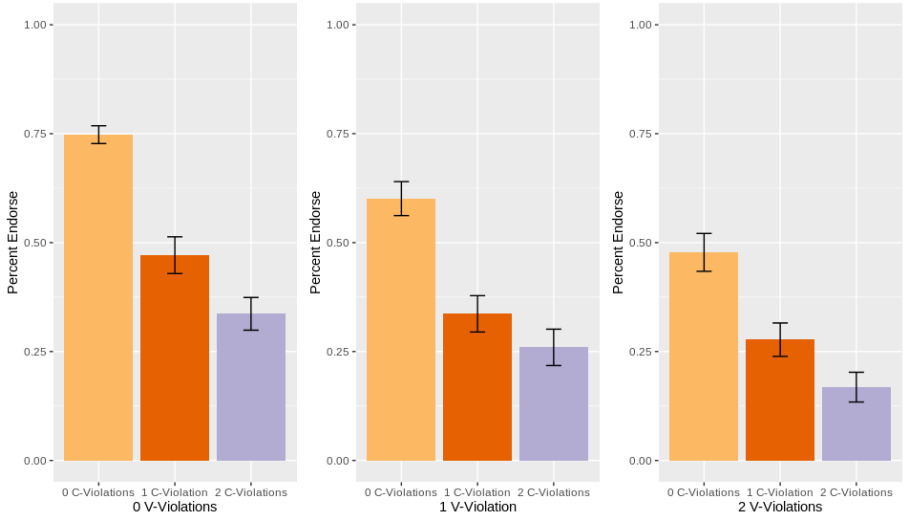

Figure 10: Results for the lexical decision task, Experiment 4. The vertical axis plots mean endorsement rate as a percentage, with standard error bars, and the horizontal axis divides the novel words according to their number of vowel-harmony violations, grouping by number of consonant violations. Note: *C-Violations* indicate violations of nasal harmony, and *V-Violations* indicate violations of backness harmony.

Figure 10 shows the results of the lexical decision task in Experiment 4. The final logistic regression model contained a random intercept for subject and word and a three-level fixed effect of violation profile. The model indicated that, holding the number of violations of nasal harmony at its average value, a nonword which violated the backness harmony phonotactic once was less likely to be endorsed than one which did not ($\beta = -1.77$, $p < 0.001$), and a form that violated the backness harmony phonotactic twice was even less likely to be endorsed than one which violated it only once ($\beta = -0.548$, $p < 0.001$). For nasal harmony the complementary was true: a single violation of nasal harmony decreased the log odds of endorsement significantly ($\beta = -0.623$, $p < 0.001$), and the addition of another violation of the same type (that is, comparing 1 vs. 2 violations) also significantly reduced the log odds

of endorsement ($\beta = -0.473$, $p = 0.002$), with the number of violations of vowel-harmony held at its average value.

6.3.2. Ratings task

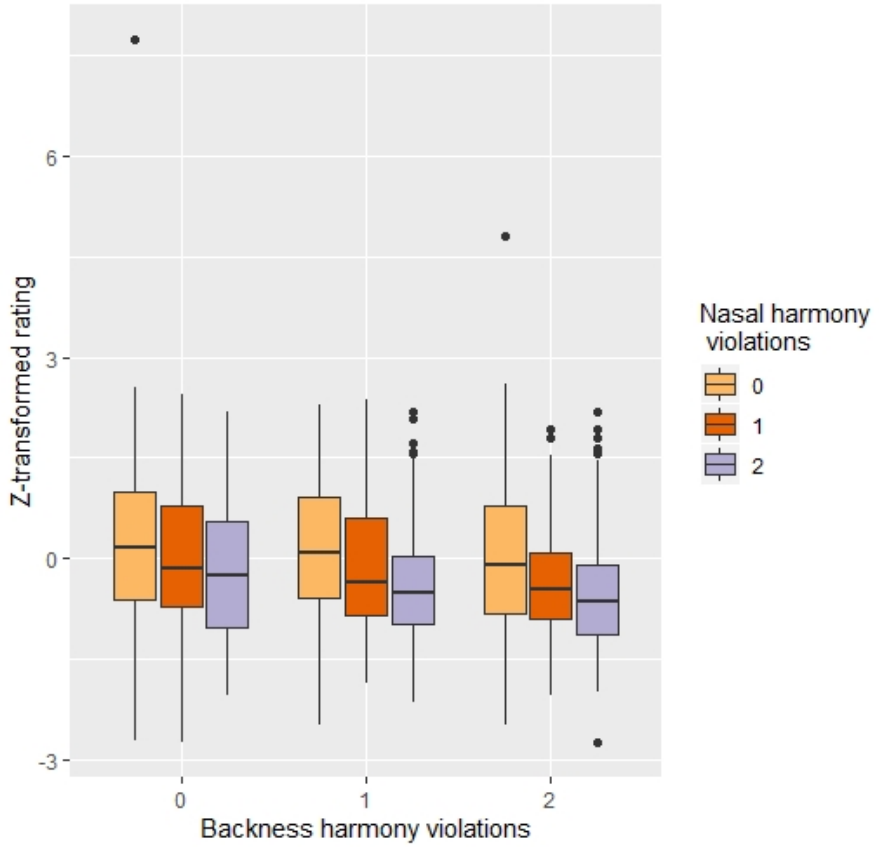

Figure 11: Results for the ratings task, Experiment 4.

Figure 11 shows the results of the ratings task in Experiment 4. The final linear regression model contained a random intercept for subject and word and two three-level fixed effect of violation profile (zero, one, or two violations). For nasal harmony, holding violations of backness harmony constant, a single violation of nasal harmony resulted in a significantly lower rating ($\beta = -6.379$, $p < 0.001$), and the further addition of a second nasal harmony violation resulted in a significant decrease in rating compared to words with only one nasal harmony violation ($\beta = -4.693$, $p = 0.002$). The model also

indicated that, holding nasal-violation level constant, a novel word which violated the backness harmony phonotactic once did not receive a significantly lower rating than one which did not ($\beta = -0.505$, $p = 0.696$), but that a form that violated the backness harmony phonotactic twice received significantly lower ratings than a word which violated it only once ($\beta = -4.439$, $p = 0.004$).

6.4. Local discussion

Experiment 4 tested for two types of constraint cumulativity, and found that learners reliably distinguish between multiple levels of well-formedness (counting cumulativity), and do so on two phonotactic dimensions simultaneously (ganging cumulativity). These findings, the one non-significant cumulative distinction in the ratings task notwithstanding, are in line with the predictions of HG theories of phonology.

Note that until this point, the definition of what constitutes a "violation" of a given phonotactic has been so simple as to have gone unsaid: in a CVCV language with consonant and vowel harmony, a word violates a phonotactic if one of the consonants or vowels doesn't match the other in the specified feature (here, [back] or [nasal]). In a language with four-syllable words, however, the definition of violation is non-trivial. For present purposes I adopt a maximally simple approach, simply counting the number of exceptions to an item-wide phonotactic pattern (i. e., the number of *disharmonic segments*). However, more linguistically sophisticated theories of vowel harmony might suppose other representational bases for phonotactic patterns: perhaps participants' knowledge of vowel harmony is better understood as being based on the number of disharmonic *pairs* of segments (cf. Bakovic [48], Pulleyblank [49], Lombardi [50]), or might distinguish between evidence based on local vs. non-local pairs (cf. Kimper [51]). Future work using this methodology could profitably distinguish between these and other proposals in the theoretical literature; for present purposes, Experiment 4 simply demonstrates that additional violations of the *same* phonotactic have cumulative effects (*counting cumulativity*) in a similar way that coincident violations of *different* phonotactics have been shown to in Experiments 1-3 (*ganging cumulativity*).

7. General discussion

This study used a series of artificial grammar learning experiments to investigate how learners acquire multiple phonotactic generalizations simultaneously, and how these generalizations interact in the grammar in a poverty-

of-the-stimulus paradigm. Experiment 1 found that learners infer ganging cumulativity among independent phonotactic violations: words violating two different phonotactic constraints were less likely to be endorsed, and received lower numerical ratings, than words which violated only one of the two. Further, this effect was linearly additive: doubly-violating words were judged ill-formed in proportion to the summed ill-formedness associated with each of their phonotactic violations. Experiment 2 replicated these findings using a different combination of phonotactics, sibilant-harmony and backness harmony, and Experiments 3a and 3b again replicated these findings using a training paradigm designed to more closely mimic natural first language acquisition. Experiment 4 used longer words to demonstrate that participants infer counting cumulativity as well, as predicted by HG. These results are significant because they demonstrate effects on phonotactic well-formedness that cannot be explained by lexical frequency asymmetries. Further, in the absence of evidence for or against constraint interaction, learners behave in ways consistent with decision heuristics employed by HG and other weighted-constraint theories of phonology, and which are explicitly predicted to not exist by OT.

Before discussing these results in light of phonological theory, it is important to address another possible explanation for the observed data. It is possible that participants were simply judging the well-formedness of novel words based on similarity to the words they heard during training, using a mechanism of analogy similar to that set forth in the Generalized Neighborhood Model (Bailey and Hahn [22]): singly-violating words differ in one location, while doubly-violating words differ in two. While quantitative comparisons between constraint-based and analogical models of well-formedness on the basis of these data are outside the scope of this paper, there are reasons to doubt that the effects observed are attributable to analogy alone, since participants exhibited both ganging and counting cumulativity when generalizing from two- to four-syllable words. Deriving these results by analogy alone would require a more complex and abstract notion of similarity than is currently thought to underpin analogical processes in grammar (see Blevins and Blevins [52] for an overview).

A repeated, yet unexpected, finding was that participants exhibited an asymmetry in the strength of markedness associated with violating different phonotactics. In Experiments 1, 3a, 3b, and 4, which paired nasal harmony with backness harmony, violations of backness harmony were judged to be less severe than violations of nasal harmony, being associated with a smaller

decrease in log odds of endorsement and ratings. However, when backness harmony was paired with sibilant harmony in Experiment 2, no such disparity was observed. What might be the source of this asymmetry?

One possibility is that the disparity is due to a by-participant effect, whereby fewer participants learned the backness harmony constraint than learned the nasal harmony constraint. I discarded this hypothesis as highly improbable, because all experiments required participants to demonstrate statistically equivalent knowledge of both phonotactics in the verification phase before being allowed to continue to the test phase.

An intriguing possibility is that the "strength" — the degree of ill-formedness incurred — of violating a given phonotactic constraint is influenced by the perceptual distinctiveness between that constraint's conforming and non-conforming instantiations. That is, the perceptual difference between a voiceless stop and a nasal stop sharing a place of articulation may be greater than the difference between a front vowel and a back vowel, or between a coronal sibilant and a post-alveolar sibilant with the same voicing specification (that is, Δ [t . . . n], [n . . . n] > Δ [o . . . e], [e . . . e], Δ [s . . . ʃ], [ʃ. . . ʃ], where Δ is a function that returns the perceptual distance between two phones). One way to formalize this notion of perceptual distance is using phonological features to measure the distance between forms. This captures the notion that nasal and voiceless homorganic stops are more perceptually distant, because they differ by two features ([nasal] and [voice]) compared to front and back vowels of the same height, as well as coronal and post-alveolar sibilants of the same voicing specification, which are less perceptually distant, differing only by one feature ([back] or [anterior], respectively).

The decreased perceptual distinctness of sibilant harmony-violations and backness harmony-violations could also be the cause for the large number of participants who were excluded from Experiment 2 for not learning both phonotactics adequately (35, in contrast to 10, 0, 0, and 0 in Experiments 1, 3a, 3b, and 4 respectively). Learning two perceptually-subtle phonotactics simultaneously may have been more difficult than learning one perceptually robust phonotactic and one perceptually subtle one, leading to increased attrition at the verification stage.

Whether the distinctiveness is better captured by phonological features or another more perceptual measure (cf. White [53], who uses confusion matrices) is beyond the scope of this paper. However, the proposed perceptually-motivated explanation is in line with theories of phonological acquisition which hold that both input statistics and perceptual similarity play a role in

shaping the grammar (Steriade [54], Wilson [55], Zuraw [56] *i. a.*).

7.1. Implications for phonological theory

This paper set out to test the predictions of two prominent phonological frameworks, HG and OT, about how learners extend multiple simultaneously-true phonotactic generalizations when faced with ambiguous training data. HG predicts that learners will exhibit both ganging cumulativity and counting cumulativity — that is, learners will be sensitive to each individual constraint violation, whether it be in the context of other violations of the same constraint, a different constraint, or alone. OT predicts that learners will attend only to the single most important constraint violation in each case, and not exhibit sensitivity to distinctions made by the presence or absence of less-important violations. The experimental results bore out HG's predictions: all experiments found that doubly-violating forms are endorsed at a significantly lower rate than singly-violating forms when the two violations come from different constraints (ganging cumulativity). Further, Experiment 4 also found evidence for counting cumulativity: words which violated a single constraint twice were judged more ill-formed than those which violated the same constraint only once (counting cumulativity), holding constant the number of violations of the other constraint. These results suggest that, if the goal of phonological frameworks is to embody salient properties of the grammars speakers possess and learn with, weighted-constraint frameworks such as HG are to be preferred over strict-ranking ones such as OT in the general case.

7.2. Why does linearity persist across tasks?

As discussed in section 2.3, finding evidence for constraint cumulativity in phonotactic learning opens the door to a range of questions about the nature of this cumulativity. The experiments presented here provide evidence that when presented with fully-conforming training data, learners infer constraint cumulativity effects that are linearly additive in nature. Further, it is important to note that this linearity is maintained across two different scales: the log odds scale in the lexical decision task, and in the 0-100 scale in the ratings task. The fact that the linearity appears to depend minimally, if at all, on the experimental task suggests that the two measures are accessing the same psychological quantity in the mind of the speaker. A linking hypothesis which can explain these facts comes from the Decathlon Model of Empirical Syntax

(Featherston [57, 58]), designed relate syntactic acceptability expressed on a continuous ratings scale to binary choice data.

In the Decathlon model, the penalty of individual items (these items are novel syntactic structures for Featherston, but the same logic applies to novel phonological items in the present context) is calculated as the dot-product of the number of violations of a given constraint and that constraint's weight (as in HG), yielding a harmony score (H). Featherston argues that it is these scores that are is accessed by continuous-scale acceptability judgments. Differences in harmony between different items under consideration are then cashed out via a candidate selection process in which items are selected (either in speech production or in a forced-choice task) in a probabilistic way such that the more well-formed candidates (those with lower H) are selected more often. While this second stage of the model was not made quantitatively explicit by Featherston, I argue that it can be modeled as a logistic regression in which differences in H between two candidates is proportional to the difference in the log odds of their being selected in a binary choice. This proposal accounts for the consistent linearity observed in cumulativity across the dependent variables in this paper: since H scores are linear combinations of penalties associated with constraint violations, the linearity at this stage propagates through further transformations depending on the experimental task, and so the linear composition of H emerged in both linear and logistic regression analyses. This hypothesis is also in line with the interpretation of constraint weights in Maximum Entropy HG ("MaxEnt"), a variant of HG which yields a probability distribution over candidates (Smolensky [59], Goldwater and Johnson [5]). MaxEnt is mathematically equivalent to multinomial regression, of which the binary logistic regression models used in to analyze the two-alternative forced-choice experimental data this paper are a more commonly-used sub-type (for more, see Jurafsky and Martin [60, ch. 5]). If this hypothesis is correct, then the present findings of linear cumulativity go beyond simply supporting HG models over OT ones, and lend support for a Decathlon-style understanding of the relationship between continuous and binary measures in acceptability judgments and for the logistic-based transformation employed in MaxEnt HG grammars more specifically. This also suggests that syntactic and phonological factors may make use of the same weighted-constraint model of well-formedness.

Stepping back from the question of why linearity should be found in two differently-scaled dependent variables, we can turn to linearity itself. While linear cumulativity may seem like an intuitive — and thus expected — re-

sult, this finding is non-trivial in light of the literature on this topic. MaxEnt HG models predict sub-linear cumulativity — individual markedness violations become less severe in the presence of others. There is some evidence for sub-linear cumulativity from Pizzo's large-scale acceptability judgment study of English nonwords. Decreasing severity for multiple coincident violations would also qualitatively align with the robust evidence for logarithmically-scaled perception of differences observed in other domains of human perception (see Fechner et al. [61] *et seq.*). There is also empirical evidence in support of super-linear cumulativity of violations. Albright [26] found that words containing two violations of relatively weak phonotactic constraints are underrepresented in the lexicon, suggesting a super-linear effect (see also Shih [62], Green and Davis [63], Smith and Pater [64] for similar findings). The specific factors which determine the linearity of cumulative interactions are left for future research.

References

[1] G. Gigerenzer, W. Gaissmaier, Heuristic decision making, Annual review of psychology 62 (2011) 451–482.

[2] A. Prince, P. Smolensky, Optimality theory: Constraint interaction in generative grammar, Optimality Theory in phonology (1993) 3.

[3] G. Legendre, Y. Miyata, P. Smolensky, Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations, Citeseer, 1990.

[4] G. Jäger, A. Rosenbach, The winner takes it allalmost: Cumulativity in grammatical variation, Linguistics 44 (2006) 937–971.

[5] S. Goldwater, M. Johnson, Learning ot constraint rankings using a maximum entropy model, in: Proceedings of the Stockholm workshop on variation within Optimality Theory, volume 111120.

[6] A. W. Coetzee, J. Pater, Lexically ranked ocp-place constraints in muna, ROA-842 (2006).

[7] J. Pater, Weighted constraints in generative linguistics, Cognitive science 33 (2009) 999–1035.

[8] K. Zuraw, B. Hayes, Intersecting constraint families: an argument for harmonic grammar, Language 93 (2017) 497–548.

[9] R. J. Scholes, Phonotactic grammaticality, volume 50, Walter de Gruyter GmbH & Co KG, 2016.

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

[11] G. Jarosz, Defying the stimulus: acquisition of complex onsets in polish, Phonology 34 (2017) 269–298.

[12] G. Guy, C. Boberg, The obligatory contour principle and sociolinguistic variation, in: Toronto Working Papers in Linguistics: Proceedings of the Canadian Linguistics Association 1994 Annual Meeting.

[13] G. R. Guy, Violable is variable: Optimality theory and linguistic variation, Language Variation and Change 9 (1997) 333–347.

[14] S. Rose, L. King, Speech error elicitation and co-occurrence restrictions in two ethiopian semitic languages, Language and Speech 50 (2007) 451–504.

[15] K. Nishimura, Lymans Law in loanwords, Ph.D. thesis, MA thesis, Nagoya University, 2003.

[16] S. Kawahara, Aspects of japanese loanword devoicing, Journal of East Asian Linguistics 20 (2011) 169–194.

[17] S. Kawahara, Japanese loanword devoicing revisited: A rating study, Natural Language & Linguistic Theory 29 (2011) 705–723.

[18] S. Kawahara, Testing japanese loanword devoicing: Addressing task effects, 2013.

[19] S. Kawahara, Lyman's law is active in loanwords and nonce words: Evidence from naturalness judgment studies, Lingua 122 (2012) 1193–1206.

[20] P. Pizzo, Investigating properties of phonotactic knowledge through web-based experimentation (2015).

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

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

[23] S. A. Frisch, N. R. Large, D. B. Pisoni, Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords, Journal of memory and language 42 (2000) 481–496.

[24] S. Shademan, Grammar and analogy in phonotactic well-formedness judgments, Ph.D. thesis, University of California, Los Angeles, 2007.

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

[26] A. Albright, Additive markedness interactions in phonology, 2012.

[27] H. Fukazawa, S. Kawahara, M. Kitahara, S.-i. Sano, Two is too much: Geminate devoicing in japanese, On-in Kenkyu 18 (2015) 3–10.

[28] S. Kawahara, S.-i. Sano, /p/-driven geminate devoicing in japanese: Corpus and experimental evidence, 2016.

[29] A. T. Martin, The evolving lexicon, Ph.D. thesis, University of California, Los Angeles Los Angeles, CA, 2007.

[30] A. Martin, Grammars leak: Modeling how phonotactic generalizations interact within the grammar, Language 87 (2011) 751–770.

[31] E. Moreton, J. Pater, Structure and substance in artificial-phonology learning, part i: Structure, Language and linguistics compass 6 (2012) 686–701.

[32] E. Moreton, J. Pater, Structure and substance in artificial-phonology learning, part ii: Substance, Language and linguistics compass 6 (2012) 702–718.

[33] J. Culbertson, Typological universals as reflections of biased learning: Evidence from artificial language learning, Language and Linguistics Compass 6 (2012) 310–329.

[34] S. Finley, Locality and harmony: Perspectives from artificial grammar learning, Language and Linguistics Compass 11 (2017) e12233.

[35] E. Moreton, Analytic bias and phonological typology, Phonology 25 (2008) 83–127.

[36] E. Glewwe, Bias in Phonotactic Learning: Experimental Studies of Phonotactic Implicationals, Ph.D. thesis, UCLA, 2019.

[37] T. Linzen, G. Gallagher, Rapid generalization in phonotactic learning, Laboratory Phonology: Journal of the Association for Laboratory Phonology 8 (2017).

[38] G. Ó. Hansson, Consonant harmony: Long-distance interactions in phonology, volume 145, Univ of California Press, 2010.

[39] R. Walker, Vowel patterns in language, volume 130, Cambridge University Press, 2011.

[40] S. Finley, Learning nonadjacent dependencies in phonology: Transparent vowels in vowel harmony, Language 91 (2015) 48.

[41] R. Lai, Learnable vs. unlearnable harmony patterns, Linguistic Inquiry 46 (2015) 425–451.

[42] M. Becker, J. Levine, Experigen–an online experiment platform, Available (April 2013) at https://github. com/tlozoot/experigen (2010).

[43] R. C. Team, et al., R: A language and environment for statistical computing (2013).

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

[45] D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal, Journal of memory and language 68 (2013) 255–278.

[46] T. Weskott, G. Fanselow, On the informativity of different measures of linguistic acceptability, Language 87 (2011) 249–273.

[47] R. A. Fisher, Two new properties of mathematical likelihood, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 144 (1934) 285–307.

[48] E. Bakovic, Harmony, dominance and control, Ph.D. thesis, 2000.

[49] D. Pulleyblank, Harmony drivers: No disagreement allowed, in: Annual Meeting of the Berkeley Linguistics Society, volume 28, pp. 249–267.

[50] L. Lombardi, Positional faithfulness and voicing assimilation in optimality theory, Natural Language & Linguistic Theory 17 (1999) 267–302.

[51] W. A. Kimper, Competing triggers: Transparency and opacity in vowel harmony, University of Massachusetts Amherst Amherst, MA, 2011.

[52] J. P. Blevins, J. Blevins, Analogy in grammar: Form and acquisition, Oxford University Press on Demand, 2009.

[53] J. White, Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a p-map bias, Language 93 (2017) 1–36.

[54] D. Steriade, The phonology of perceptibility effects: the p-map and its consequences for constraint organization, Ms., UCLA (2001).

[55] C. Wilson, Learning phonology with substantive bias: An experimental and computational study of velar palatalization, Cognitive science 30 (2006) 945–982.

[56] K. Zuraw, 1* map constraints (2013).

[57] S. Featherston, The decathlon model of empirical syntax, Linguistic evidence: Empirical, theoretical, and computational perspectives (2005) 187–208.

[58] S. Featherston, 6 the decathlon model, Current Approaches to Syntax: A Comparative Handbook 3 (2019) 155.

[59] P. Smolensky, Information processing in dynamical systems: Foundations of harmony theory, Technical Report, Colorado Univ at Boulder Dept of Computer Science, 1986.

[60] D. Jurafsky, J. H. Martin, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 3rd Edition Draft, Pearson/Prentice Hall Upper Saddle River, 2019.

[61] G. T. Fechner, D. H. Howes, E. G. Boring, Elements of psychophysics, volume 1, Holt, Rinehart and Winston New York, 1966.

[62] S. S. Shih, Constraint conjunction in weighted probabilistic grammar, Phonology 34 (2017) 243–268.

[63] C. R. Green, S. Davis, Superadditivity and limitations on syllable complexity in bambara words, Perspectives on phonological theory and development, in honor of Daniel A. Dinnsen (2014) 223–47.

[64] B. W. Smith, J. Pater, French schwa and gradient cumulativity, Manuscript (University of California, Berkeley and University of Massachusetts, Amherst) (2017).