# Constraint cumulativity in phonotactics: evidence from Artificial Grammar Learning studies[⋆]

Canaan Breiss

*University of California, Los Angeles*

## Abstract

An ongoing debate in phonology concerns the treatment of *cumulative constraint interactions*, or "gang effects", which in turn bears on the question of which phonological frameworks are suitable models of the grammar. This paper uses a series of Artificial Grammar Learning experiments to examine the inferences learners draw about such cumulative constraint violations in phonotactics using a poverty-of-the-stimulus design. I show that learners consistently infer linear *counting* and *ganging* cumulativity between a range of phonotactic violations. I evaluate three phonological frameworks — the classical Recursive Constraint Demotion algorithm (Tesar and Smolensky, 2000), Stochastic Optimality Theory (Boersma and Hayes, 2001) and Maximum Entropy Harmonic Grammar (Smolensky, 1986; Goldwater and Johnson, 2003) — on their ability to model such cumulativity when exposed to the same training data as the experimental subjects. I find that while the Stochastic Optimality Theory and Maximum Entropy frameworks are able to capture the ganging cumulativity participants displayed, only the

[⋆]Many thanks to Bruce Hayes and Megha Sundara for their guidance and advice on all aspects of this research. Thanks also to Adam Albright, Adam Chong, Karthik Durvasula, Sara Finley, Shigeto Kawahara, James White, Colin Wilson, two anonymous reviewers for Phonology, and members the audience at the LSA Annual Meeting 2019, SCAMP 2019, and the UCLA Phonology seminar for much helpful commentary and discussion. I am also grateful to 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. Supplementary materials can be accessed on the Phonology journal website, and on the author's homepage at PUT FINAL LINK. Full responsibility for all remaining errors rests with the author. This work was supported by NSF Graduate Research Fellowship DGE-1650604 to Canaan Breiss.

Maximum Entropy model captures both counting *and* ganging cumulativity. This follows directly from MaxEnt's use of weighted, rather than ranked, constraints.

*Keywords:* phonotactics, cumulative constraint interaction, gang effects, poverty of the stimulus, artificial grammar, counting cumulativity, ganging cumulativity

## 1. Introduction

.

The treatment of *cumulative constraint interactions* is the subject of ongoing debate in phonology. In this paper I take on the topic of cumulative constraint interactions in phonotactics. I focus on the phonotactic domain because it allows for a direct interrogation of the relationship between cumulative constraint violations and acceptability. This method is inspired by work in the field of experimental syntax (Featherston, 2005, 2019), where syntactic violations are manipulated in a crossed experimental design to tease apart the independent contribution of each violation, and gain insight into how multiple violations are combined in the grammar. In the first part of the paper, I combine this independent manipulation of violations with an Artificial Grammar Learning (AGL) paradigm which imposes a poverty-of-the-stimulus learning environment. By doing so I ensure that whatever generalizations participants form about the (non)interaction of independent phonotactic violations can be taken to reflect properties of the structure of the grammar, rather than artefacts of language-specific distributional factors.

In the second part of the paper, I use the experimental results to evaluate three phonological frameworks — the Recursive Constraint Demotion algorithmTesar and Smolensky (2000), Stochastic OT (Boersma and Hayes, 2001) and Maximum Entropy HG (Smolensky, 1986; Goldwater and Johnson, 2003) — on their ability to model the observed cumulativity.

## 2. Constraint cumulativity in phonological theory

Constraint-based phonological frameworks diverge on whether they can model cumulative constraint interactions. Classical Optimality Theory ("OT"; Prince and Smolensky (1993), *et seq.*) holds that speakers are informationally-frugal when computing phonological well-formedness: constraints on well-formed structures are strictly ranked, and the choice between possible outcomes is determined by the highest-ranking constraint that distinguishes between them. By contrast, Harmonic Grammar ("HG"; Legendre et al. (1990), *et seq.*) holds that speakers take an informationally-holistic approach, considering all constraint violations when choosing the optimal outcome. The difference can be observed in the schematic tableaux below. In OT, candidate B wins out at the expense of candidate A, because candidate A violates the higher-ranked Constraint 1, while candidate B does not. Because Constraint 1 is ranked above Constraint 2, 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, and candidate B is deemed optimal.

| | Constraint 1 | Constraint 2 |
|---|---|---|
| *Candidate A* | *! | |
| ☞*Candidate B* | | ** |

Table 1: Schematic example of an OT tableau.

In HG the optimal outcome is the one which has the lowest harmony penalty ($H$) when considering *all* violations. Each candidate's harmony is equal to the number of times it violates each constraint, multiplied by the weight of the constraint violated. Using this method, the same violations result in candidate A being optimal because it has a lower harmony than candidate B.[1] This is because the two violations of Constraint 2, though tolerated individually, together outweigh the penalty associated with the single violation of Constraint 1.

| | Constraint 1 | Constraint 2 | |
|---|---|---|---|
| *weights:* | 3 | 2 | $\mathscr{H}$ |
| ☞*Candidate A* | * | | 3 |
| *Candidate B* | | ** | 4 |

Figure 1: Schematic example of a HG tableau.

[1] Of course, this only holds when the specific weights of the constraints involved permit it; the weights are chosen in this schematic example to mirror dominance relations in OT for the sake of demonstration.

HG and OT predict different outcomes from the same same schematic example because candidates' violations are cumulative in HG but not in OT.[2]

There are two possible types of constraint cumulativity (Jäger and Rosenbach, 2006): *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 or more violations of the first constraint are together more penalizing than the single violation of the second. Ganging cumulativity occurs when 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 yield a more severe penalty.

## 3. Constraint cumulativity in phonological typology

### *3.1. Constraint cumulativity in alternations*

Evidence for constraint cumulativity is often discussed in the context of conditions on the relationship between phonological inputs and outputs (Goldwater and Johnson, 2003; Coetzee and Pater, 2006; Pater, 2009; Zuraw and Hayes, 2017, *i. a.*). Within cumulativity, the majority of cases are of *ganging cumulativity*, where two (or more) distinct factors interact. For instance, the cumulative combination of factors influencing the likelihood of *-t/-d* deletion in corpus data was noted by Guy and Boberg as early as 1994, and the difficulty the phenomenon presented for OT was pointed out not long after (Guy, 1997). More recently, Rose and King (2007) used a speech-error elicitation task to examine 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 several constraints at once than when stimuli violated each constraint independently. Pater (2009) analyzes data from Japanese loan words (originally from Nishimura (2003)) to argue that a static phonotactic restriction known as "Lyman's Law" which prohibits multiple voiced obstruents within a word can be construed as a case of constraint cumulativ-

[2] Technically, in OT markedness violations can be compared numerically when two candidates tie on all higher-ranked constraints; these cases are not typically considered among discussions of cumulative constraint interaction, however.

ity.[3] 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 them, enforcing the upper limit on voiced obstruents. Kawahara (2011a,b, 2013) follows this formal analysis up with a series of acceptability judgment studies, finding robust support for Pater's conclusions. Kawahara also finds experimental evidence that Lyman's Law violations can block a process of inter-morphemic obstruent voicing known as *rendaku* (Kawahara, 2012) in a further case of apparent cumulativity, supporting the observations made in Itô and Mester (1986). Recent studies by Kawahara (*submitted*); Kawahara and Breiss (*submitted*) also indicate that the relationship between form and meaning characteristic of sound-symbolism displays cumulative effects. On the *counting cumulativity* front there is less work, though recent findings by Kim (2019); Kumagai (2017) demonstrate that only two nasals (but not one) blocks *rendaku* application in Japanese compounds.

### *3.2. Phonotactics as a testing ground for theories of cumulativity*

Constraint cumulativity has also been demonstrated in phonotactics, generally taking the form of additive effects of multiple markedness violations on the likelihood of lexical attestation or experimentally-assessed acceptability. Evidence comes from the study by Pizzo (2015), mentioned above, as well as that of Albright (2009), which contains similar findings. Taken at face value, these studies seem to constitute strong evidence for the cumulativity of markedness constraints — multiple simultaneously-violated constraints together have an effect on speakers' judgments which is greater than that of each constraint alone.

### *3.3. The lexicon as a confound in the study of the phonotactic grammar*

While suggestive of cumulative behavior, however, such findings have an alternative explanation. This is because in each of these cases, experimentally-determined well-formedness is highly correlated with the frequency of such structures are in the lexicon. Even setting aside models which explicitly use the number of similar words in the lexicon to estimate acceptability (ex., the Generalized Neighborhood Model of Bailey and Hahn (2001)), the

[3] A reviewer notes that, depending on one's theoretical orientation, loanword adaptation might be construed as a phonotactic repair, rather than a phonological alternation.

prominent role of lexical statistics in influencing well-formedness judgments is well established in generative phonology. Pioneering work by Coleman and Pierrehumbert (1997) highlighted the connections between the lexicon and phonotactic well-formedness in their predictive model of nonword judgments, inspiring much further work by Frisch et al. (2000); Shademan (2007); Daland et al. (2011); Jarosz and Rysling (2017), among many others. Albright (2012); Fukazawa et al. (2015) and Kawahara and Sano (2016) 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 (2007, 2011) found that, assuming speakers prefer to reuse novel coinages which are phonotactically well-formed, the lexicon can come to underrepresent phonotactically ill-formed words over time. This sets 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 in the context of the lexicon, or whether skewed lexical statistics are the product of the phonological grammar — cannot be satisfactorily resolved. This prevents us from taking evidence of phonotactic cumulativity in natural languages such as Albright (2012); Pizzo (2015) as evidence of the nature of the grammar which is unbiased by lexical statistics.

## 4. Experimental design

To deconfound phonotactic acceptability and lexical frequency, I used an AGL paradigm to create a "sandbox environment" where these factors could be carefully controlled. This allows me to interpret participants' inferences about such (non)cumulativity, made in the absence of disambiguating evidence and distributional asymmetries of the lexicon, to be revealing of the nature of phonotactic grammar, and not simply a case of "frequency-matching" (cf. Ernestus and Baayen (2003)) in acceptability judgments.

Turning to the specific phonotactics involved, all experiments in this paper paired varieties of consonant and vowel harmony. These phenomena have traditionally constituted core objects of generative phonological analysis (see Hansson (2010) and Walker (2011) for overviews of consonant and vowel harmony patterns respectively), and both have been successfully learned in other AGL experiments (ex., Finley (2015); Lai (2015)). Because the 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.[4] In Experiments 1, 3a-b, and 4 I used consonant nasality harmony (hereafter *nasal harmony*): a word's consonants agree in nasality, being either drawn from the set of nasal stops {/m, n/} or voiceless oral stops {/p, t/}; for a survey of parallels in natural languages, see Hansson (2010, p. 111 *et seq.*). In Experiment 2 I used *sibilant harmony*: sibilant consonants in a word agree in anteriority, being drawn from {/s, z/} or {/ʃ, ʒ/} (see Hansson (2010, p. 55 *et seq.*) for a typological survey). All experiments used vowel backness (as well as rounding) harmony — referred to hereafter as simply *backness harmony*: all vowels in a word agreed in backness, being drawn from the set of {/i, e/} or {/u, o/} (for an overview, see Walker (2011)).

To ensure an accurate assessment of participants' well-formedness judgments, I elicited acceptability judgments from participants using two different tasks. First, participants rendered binary discrimination judgements in what I term the *lexical decision task*, in which they were asked to judge whether a novel word 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 that word sounded as an example of the language they had learned. Robust support of either outcome — whether speakers display cumulativity or not — should be the result of converging evidence from both dependent variables.[5]

[4] An overlap in domain of the phonotactics could lead to potentially complex interactions between the two in learning. For example, if one phonotactic regulated labiality of consonants and vowels and another regulated height harmony among vowels, the overlap of these harmonies could lead participants to draw spurious conclusions about the height and rounding features of vowels specifically, subject to the inter-tier bias discussed in Moreton (2008), in addition to or instead of one of the intended phonotactic patterns.

[5] A reviewer raises the possibility that participants could simply be judging the well-

## 5. Experiment 1: Ganging cumulativity, nasal and backness harmonies

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

### *5.1. Methods*

#### *5.1.1. 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.

#### *5.1.2. 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

formedness of novel items based on some non-phonological measure of similarity (such as $n$-gram probabilities of the string) based on the items seen in training phase, and thus not be inducing markedness constraints at all. While this is theoretically possible, since all phoneme bigrams in the generalization items appeared in the exposure items, such a generalization would come down to either tracking tier-based $n$-grams, or else tracking counts over trigram windows of the string. The degree to which these generalizations are “non-linguistic” is debatable, however, and a topic of ongoing investigation (cf., e.g., Wilson and Gallagher (2018)). Here I proceed on the assumption that whatever types of generalizations participants are forming are at least linguistically-informed and thus are in the domain of the two generative theories I test here, but leave open the exact structure of these generalizations underpsecified for the present paper (though see Durvasula and Liter (2020) for recent work focusing on at exactly what level of representational granularity learners form generalizations in AGL experiments). Interested readers can access the raw data in the supplementary materials.

by changing one of the consonants or vowels in the fully-conforming word. Thus the pair of words differed only in a single instance of that segment. 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 *potu* vs. *ponu*.

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 (in line with the general findings of Gallagher (2014)). 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.

### *5.1.3. 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 in which words were presented auditorily without feedback. 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, which were not accompanied by feedback about accuracy. On each trial, participants were asked to choose which of the two words belonged to the language they had learned in training; if participants scored above 80% (13 or more correct answers out of 16 trials) in the verification phase 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 error 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 to the 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 they sounded as an example of the language they had learned. 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.

#### *5.1.4. Procedure*

Participants were tested individually in a sound-attenuated room using a modified version of the Experigen platform (Becker and Levine, 2010). After giving their informed consent to participate in the study, the experiment began with participants being 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 (via on-screen text) 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 also had a similar structure, with each task consisting of a series of trials containing one word which participants were instructed to repeat out loud before either making the lexical decision or assigning it a rating.

### *5.2. Analysis*

Data from the test phase was analyzed using mixed effects regression models in R (R Core Team, 2013) using the *lme4* package (Bates et al., 2015). All statistical analyses began by fitting a maximally-specified model (following Barr et al. (2013)), which contained a random intercept for subject and item, fixed effects of violation of backness harmony, violation of nasal harmony, and their interaction, and random slopes for all fixed effects by subject. Dummy coding was used for the two fixed effects. 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 modeled the raw numerical data using a linear model. Planned comparisons were was used to test for a difference between the singly-violating levels and the doubly violating level using the *glht()* function from the `multcomp` package (Hothorn et al., 2016); this is done to verify that the difference between singly-violating and doubly-violating levels was significant, a more rigorous check for cumulativity which complements a lack of observed significant interaction. If the main effects of phonotactic violation are significant, and the interaction of the two is not, we can conclude that participants learned a penalty for violating each constraint independently, and we cannot conclude they inferred anything other than linear cumulativity between the phonotactics. Confirmation that both singly-violating forms were judged significantly better than the doubly-violating forms confirms that the interaction was not sub-linear; a lack of significant difference on this score does not confirm a lack of sub-linearity, but neither does it provide evidence in favor of it, since the interaction between phonotactic violations was not significant. Note that although I regressed on the raw ratings data, for the sake of legibility I plot $z$-normalized ratings throughout the paper.

### *5.3. Results*

#### *5.3.1. Lexical decision task*

Figure 2 shows the results of the lexical decision task in Experiment 1.

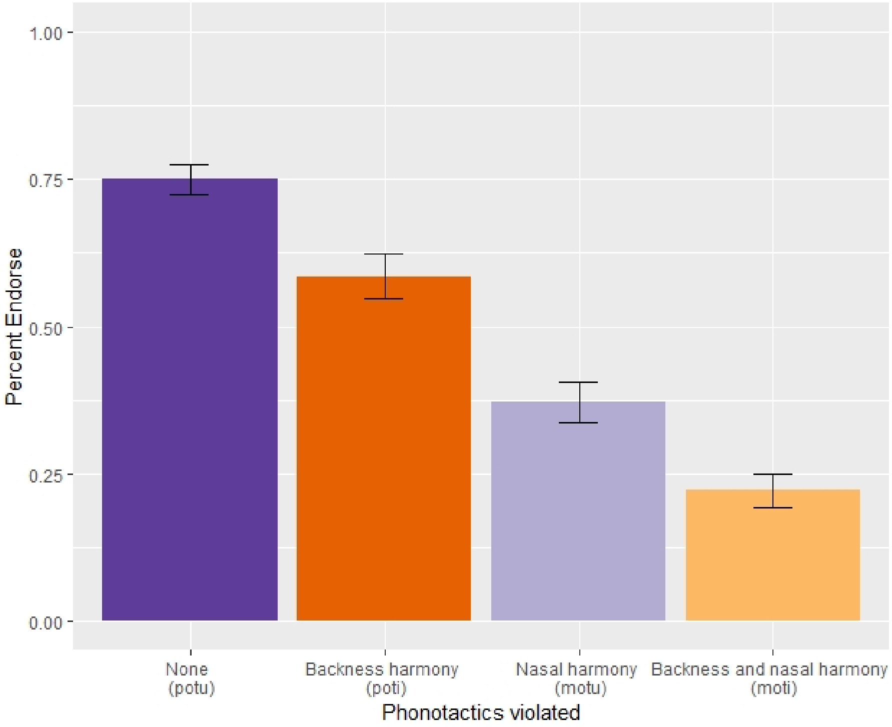

Figure 2: Results for the lexical decision task, Experiment 1. The vertical axis plots mean endorsement rate — the likelihood of an individual item of a given profile being judged as being able to be a part of the language in question — as a percentage with standard error bars, and the horizontal axis divides the novel words according to their phonotactic violation profile, together with an illustrative example of that profile type.

The final model contained a random intercept for subject and word. There was a main effect of violating backness harmony ($\beta = -0.813$, std. err. $= 0.201$, $z = -4.044$, $p < 0.001$) and a main effect of violating nasal harmony

($\beta$ = $-1.748$, std. err. = 0.202, $z$ = $-8.646$, $p$ < 0.001). The interaction between the two was not significant ($\beta$ = 0.020, std. err. = 0.321, $z$ = 0.061, $p$ = 0.951) . This means that forms violating backness harmony were less likely to be endorsed than forms that did not, and forms that violated nasal harmony were less likely to be endorsed than those that did not. Further, there is no evidence to think that doubly-violating forms were not endorsed at a rate proportional to the summed penalty for each of their violations. The post-hoc comparisons indicated that forms violating only nasal harmony trended towards a significant difference in log-odds of endorsement from doubly-violating forms ($\beta$ = 0.833, std. err. = 0.474, $z$ = 1.758, $p$ = 0.079) , and that forms violating only backness harmony differed significantly from doubly-violating forms ($\beta$ = 1.768, std. err. = 0.475, $z$ = 3.723, $p$ < 0.001).[6]

### *5.3.2. Ratings task*

Results of the ratings task are presented in Figure 3. The model for the ratings task had the same random effect structure as that of the lexical decision task. There was a main effect of violating backness harmony ($\beta$ = $-7.1$, std. err. = 2.918, $t$ = $-2.433$, $p$ = 0.020), which yielded a decrease in ratings. There was also a main effect of violating nasal harmony ($\beta$ = $-23.676$, std. err. = 2.918, $t$ = $-8.115$, $p$ < 0.001); their interaction was not significant ($\beta$ = $-2.025$, std. err. = 4.614, $t$ = $-0.439$, $p$ = 0.663). Post-hoc comparisons indicated that forms violating only nasal harmony did not significantly differ in rating from those violating both backness and nasal harmony ($\beta$ = 5.075, std. err. = 6.843, $z$ = 0.742, $p$ < 0.458), but forms violating only backness harmony did differ from doubly-violating forms ($\beta$ = 21.651, std. err. = 6.843, $z$ = 3.164, $p$ = 0.002).

### *5.4. Local discussion*

Experiment 1 provides evidence that in a poverty-of-the-stimulus environment — that is, in a context unbiased by lexical statistics — learners infer ganging cumulativity between violations of two separate phonotactic constraints in their learning data. In the lexical decision task, doubly-violating

[6] An identical model with a random slope of *presentation* (*first* vs. *second*) by item was also fit, to see whether each item being seen more than once affected the results; the findings are qualitatively unchanged, and quantitatively extremely close to those of the model reported here.

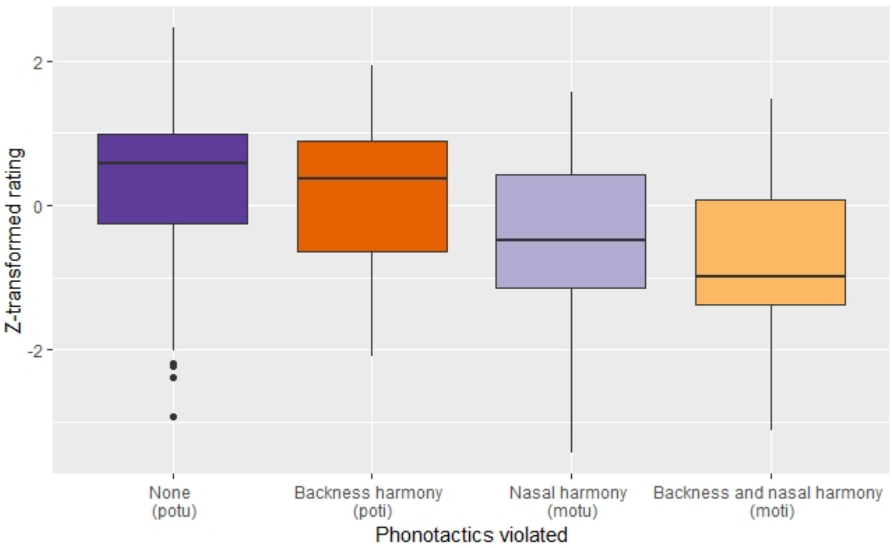

Figure 3: Results for the ratings task, Experiment 1. For the remainder of the paper, the central line in each boxplot indicates the median, with the colored portion of the box extending from the 25th percentile to the 75th; whiskers extend a further 1.5 times the inter-quartile range of the data. For 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.

forms were endorsed in proportion to the likelihood of endorsement of forms bearing each of their violations independently, while in the ratings task there was a main effect of violating both phonotactics, but it is unclear whether the cumulativity was linear or sub-linear (that is, whether only nasal harmony-violating forms differed from doubly-violating forms). In either case, however, the overall finding is that cumulativity obtains.

## 6. Experiment 2: Ganging cumulativity, sibilant and backness harmonies

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

### *6.1. Methods*

#### *6.1.1. 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. Note that although the structure of verification structure for this experiment was identical to that of Experiment 1, this experiment had an extremely high attrition rate, about 50%; in section 9 I consider what might have been the cause of this dramatic difference.[7] Recruitment method, compensation, experimental setting, and software were the same as for Experiment 1. New materials were created for Experiment 2 by replacing /p/ with /ʃ/, /m/ with /ʒ/, /t/ with /s/, and /n/ with /z/. 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.

### *6.2. Results*

#### *6.2.1. Lexical decision task*

Figure 4 shows the results of the lexical decision task in Experiment 2. The final logistic regression model contained a random intercept for subject and word. There was a main effect of violating backness harmony ($\beta = -1.223$, std. err. $= 0.177$, $z = -6.992$, $p < 0.001$) and also a main effect of violating sibilant harmony ($\beta = -0.968$, std. err. $= 0.176$, $z = -5.497$, $p < 0.001$); the interaction of these factors was not significant ($\beta = 0.283$, std. err. $= 0.281$, $z = -1.007$, $p = 0.314$). The post-hoc comparison between only backness harmony-violating forms and doubly-violating forms was significant ($\beta = 1.506$, std. err. $= 0.415$, $z = 3.628$, $p < 0.001$), as was the comparison between only sibilant harmony-violating forms and doubly-violating forms ($\beta = 1.251$, std. err. $= 0.414$, $z = 3.018$, $p = 0.003$).

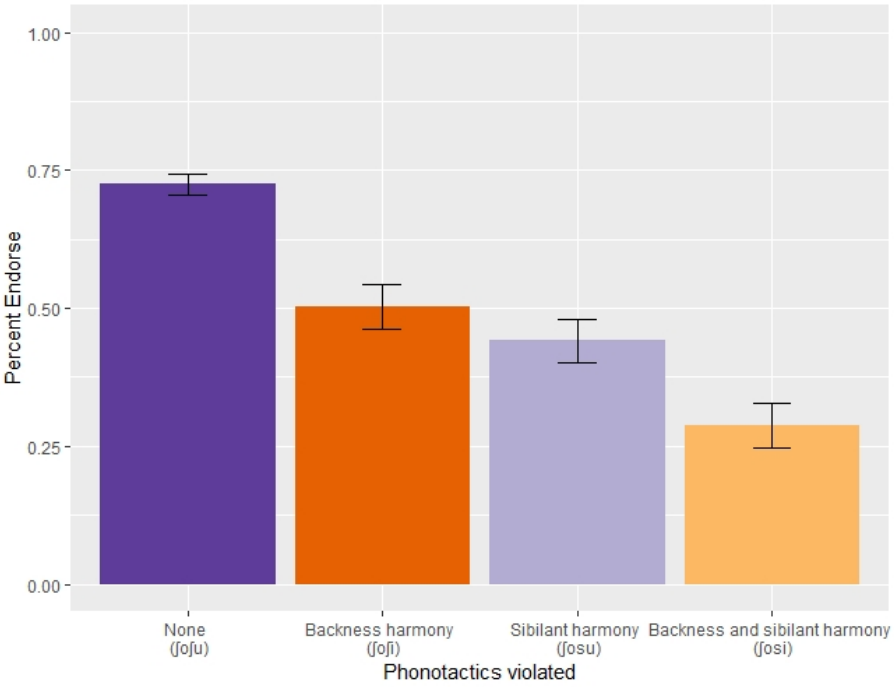

Figure 4: Results for the lexical decision task, Experiment 2.

### *6.2.2. Ratings task*

Results of the ratings task are presented in Figure 5. 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, violation of backness harmony resulted in significantly lower ratings ($\beta = -14.429$, std. err. $= 2.337$, $z = -6.070$, $p < 0.001$), as did violations of sibilant harmony ($\beta = -14.722$, std. err. $= 2.374$, $t = -6.200$, $p < 0.001$); the interaction of these factors was not significant ($\beta = -1.008$, std. err. $= 3.756$, $z = -0.269$, $p = 0.79$). Post-hoc comparisons revealed that the difference in rating between only backness harmony-violating forms and

[7] A reviewer raises the possibility that the high attrition rate may have lead to the results of the experiment being systematically biased based on the pool of participants from whom data was collected. Although this is a theoretical possibility, I judge it unlikely that a failure to learn individual phonotactics might impact the way these phonotactics — when learned successfully — interact cumulatively in the grammar.

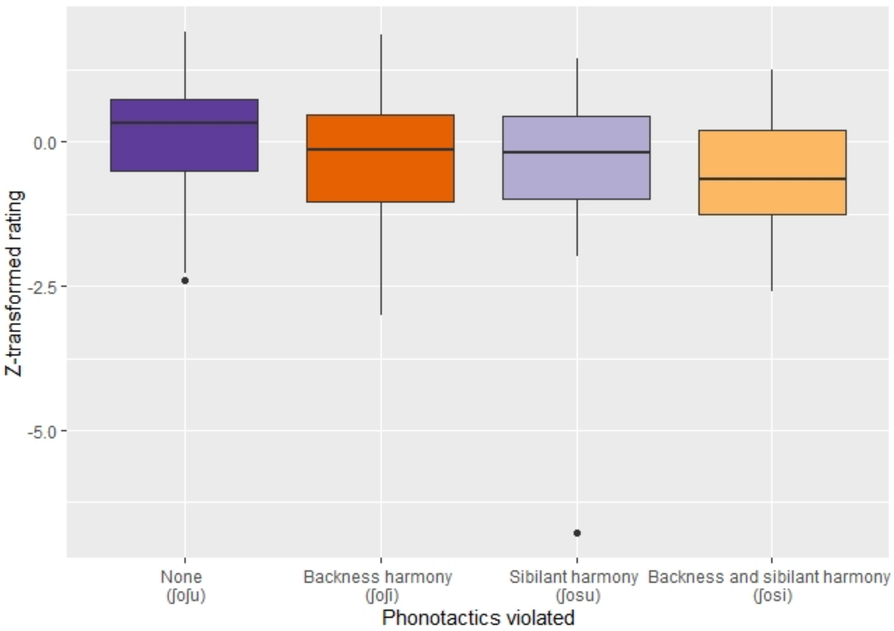

Figure 5: Results for the ratings task, Experiment 2.

doubly-violating forms was significant ($\beta = 13.421$, std. err. $= 5.573$, $z = 2.408$, $p = 0.016$), as was the difference between only sibilant-violating forms and doubly-violating forms ($\beta = 13.713$, std. err. $= 5.569$, $z = 2.462$, $p = 0.014$).

### *6.3. Local discussion*

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

## 7. Experiment 3a-b: Passive learning of ganging cumulativity, nasal and backness harmonies

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

### *7.1. Methods*

### *7.1.1. Participants*

76 new undergraduate students were recruited to participate in this experiment. 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.

### *7.1.2. 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. Since the exposure was designed to be naturalistic, I did not impose an absolute threshold for advancement to the test phase; instead participants were allowed to advance to the test phase if they did not make significantly more errors on verification trials which contrasted in vowel harmony violation only compared to those those which contrasted in consonant harmony violation only, and vice versa.[8] 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.

### *7.1.3. 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

[8] I used Fisher's exact test (Fisher, 1934) to determine the level at which the proportion of correct answers for each phonotactic significantly differed, across the range of possible accuracies. The maximum difference between the number of errors participants could make on each type of verification trial without being significantly different by this measure was 3.

sit and listen to the speech stream. 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.

### *7.2. Results I: Do subjects infer ganging cumulativity?*

#### *7.2.1. Lexical decision task*

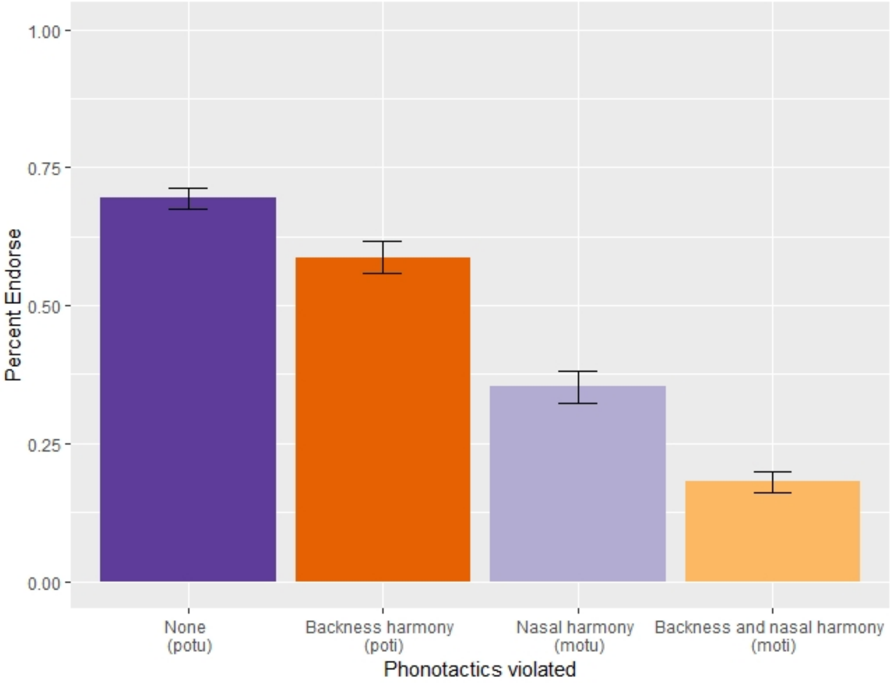

Figure 6: Results for the lexical decision task, Experiment 3a.

Figure 6 shows the results of the lexical decision task in Experiment 3a. The final logistic regression model contained a random intercept for subject and word. Violating backness harmony was associated with a significant decrease in log-odds of endorsement ($\beta = -0.504$, std. err. $= 0.226$, $z = -2.233$, $p = 0.0.026$), as was was violating nasal harmony ($\beta = -1.562$, std. err. $= 2.227$, $z = -6.900$, $p < 0.001$); the interaction between the two was not significant ($\beta = -0.486$, std. err. $= 0.363$, $z = -1.292$, $p = 0.196$). Post-hoc comparisons revealed that forms only violating backness harmony were significantly more likely to be endorsed than doubly-violating forms ($\beta$

$= 1.094$, std. err. $= 0.533$, $z = 2.049$, $p = 0.041$), while forms violating only nasal harmony were not significantly more likely to be endorsed than doubly-violating forms ($\beta = 0.034$, std. err. $= 0.533$, $z = 0.065$, $p = 0.948$).

### *7.2.2. Ratings task*

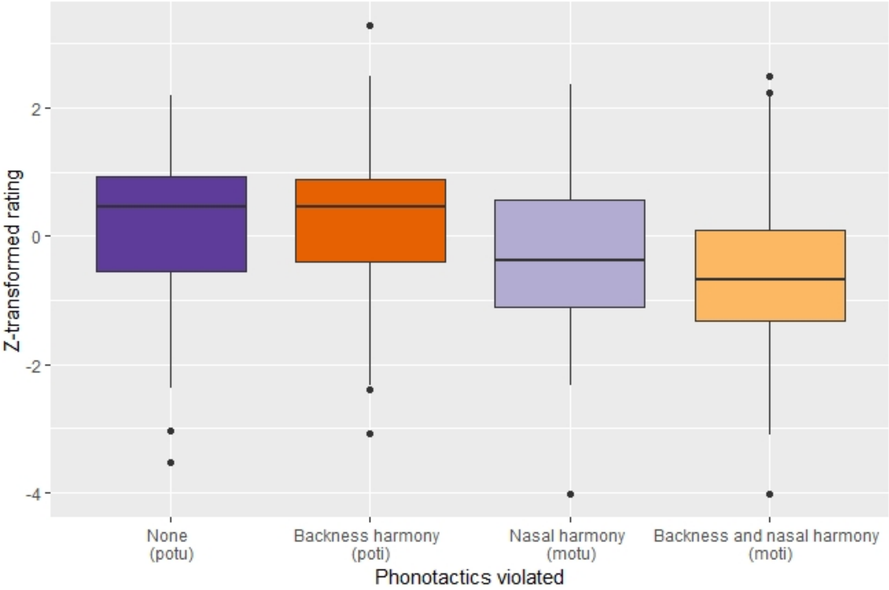

Figure 7: Results for the ratings task, Experiment 3a.

Results of the ratings task are presented in Figure 7. The main effect of violating nasal harmony was significant ($\beta = -17.536$, std. err. $= 3.707$, $z = -4.730$, $p < 0.001$), but the main effect of violating backness harmony was not ($\beta = 0.706$, std. err. $= 3.706$, $z = 0.190$, $p = 0.850$), nor was the interaction between these factors ($\beta = -8.666$, std. err. $= 5.861$, $z = -1.478$, $p = 0.146$). Since the model did not indicate that there was a main effect of violating backness harmony, I did not conduct the post-hoc tests.

### *7.3. Local discussion*

Experiment 3a provides some evidence for the robustness of inferred cumulativity under more naturalistic passive exposure training: in the lexical decision task, violations of both the nasal harmony and backness harmony phonotactics contributed independently to likelihood of endorsement. In the ratings task, however, only violations of the nasal harmony phonotactic contributed to lower ratings on average. Experiment 3b sought to replicate the null effect observed in the ratings task.

## Experiment 3b

In an attempt to replicate the null effect of cumulativity observed in the ratings task in Experiment 3a, a shortened, ratings-only version of the same experiment was carried out. If the results of the ratings task in Experiment 3a were simply the result of random fluctuation in the experimental outcome, we expect to observe cumulativity in Experiment 3b. If, on the other hand, this difference should be attributed to substantive differences between the designs of Experiments 1-2 and 3a, we should expect to replicate the null result .

### *7.4. 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, software, materials, and analysis were the same as for Experiment 1, except that participants did not complete the lexical decision task during the test phase.

### *7.5. Results I: Do subjects infer ganging cumulativity?*

Results of the ratings task are presented in Figure 8. Mirroring results of the lexical decision task from Experiment 3a, there was a main effect of violating backness harmony ($\beta = -10.136$, std. err. = 3.875, $z = -2.615$, $p = 0.012$), and a main effect of violating nasal harmony ($\beta = -27.029$, std. err. = 3.875, $z = -6.974$, $p < 0.001$); the interaction between the two was not significant ($\beta = 0.304$, std. err. = 6.129, $z = -0.050$, $p = 0.961$). Post-hoc comparisons indicated that forms violating only backness harmony significantly differed in rating from doubly-violating forms ($\beta = 27.332$, std. err. = 9.089, $z = -3.007$, $p = 0.003$), and forms which violated only nasal harmony did not differ significantly from doubly-violating forms ($\beta = 10.439$, std. err. = 9.089, $z = 1.149$, $p = 0.251$).

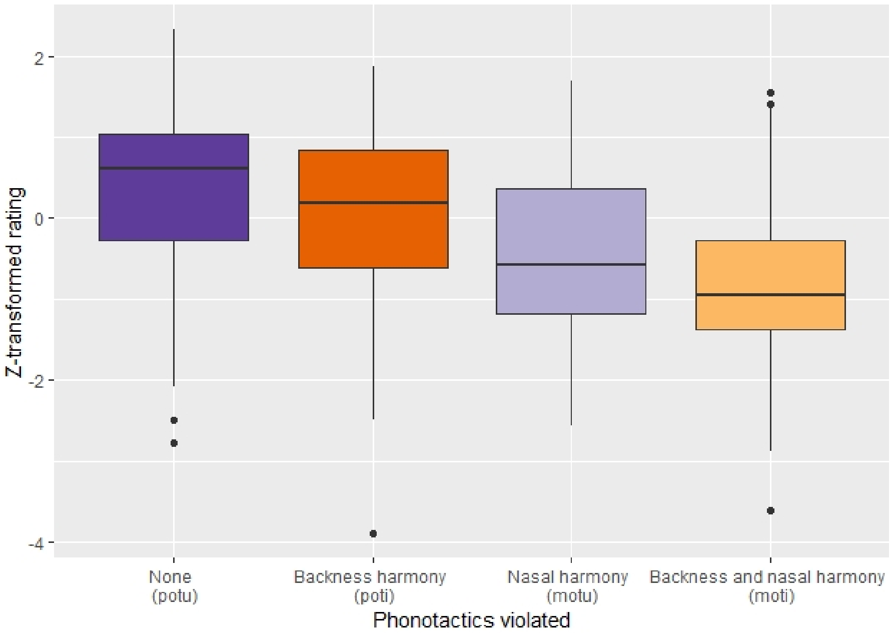

Figure 8: Results for the ratings task, Experiment 3b.

### *7.6. Local discussion*

Experiment 3b fails to replicate the null effect observed in the ratings task in Experiment 3a.

## 8. Experiment 4: Ganging and counting cumulativity

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, predicts that a word violating the a constraint $n$ times will be less well-formed than a near-identical word violating the same constraint $n-1$ times. To test this prediction, participants were tested on longer novel words which allowed for each word to host up to two violations of each phonotactic constraint. This also allowed me to test whether counting and ganging cumulativity obtained simultaneously, since I examined a number of violations of each single phonotactic in the context of each level of the other, yielding a fully-crossed design.

### *8.1. Methods*

#### *8.1.1. Participants*

71 new undergraduate students were recruited to participate in this experiment. 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.

#### *8.1.2. Materials*

Training and verification materials were identical to those of 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 locus of violation vs. two) of both phonotactics.[9]

#### *8.1.3. 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 (one run through all 48 forms in generalization per task, rather than two), 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 licit words of many lengths.[10]

[9] I delay the formalization of these phonotactics until section 11.1.

[10] A reviewer expressed concern that this overt mention of word-length biased participants by encouraging them to treat the words in the generalization phase differently than they might otherwise. Although technically a possibility, I did not judge this to be a likely source of systematic bias in the experiment, since participants saw *only* four-syllable words in the generalization phase. It is possible that without this direction participants would have given the four-syllable words lower ratings, or endorsed them at a lower rate, across the board. However, I judged this to not be a worrying possibility, because across-the-board effects in acceptability should not impact any cumulativity that their judgements displayed.

### *8.2. 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. Rather than run tests to determine whether each of the nine categories of stimulus in the generalization phase were significantly different from one another, I compared a model that contained a two binary fixed effects (whether or not a form violated backness harmony (*yes* / *no*, whether or not a form violated nasal harmony (*yes* / *no*) to a model which contained two corresponding three-level factors denoting *how many* times a form violated each phonotactic. Though a less rigorous statistical test, used here because of a lack of experimental power, it directly corresponds to the question that is at stake in linguistic theory: does a model which allows counting cumulativity capture the data better than one which does not?

### *8.3. Results*

#### *8.3.1. Lexical decision task*

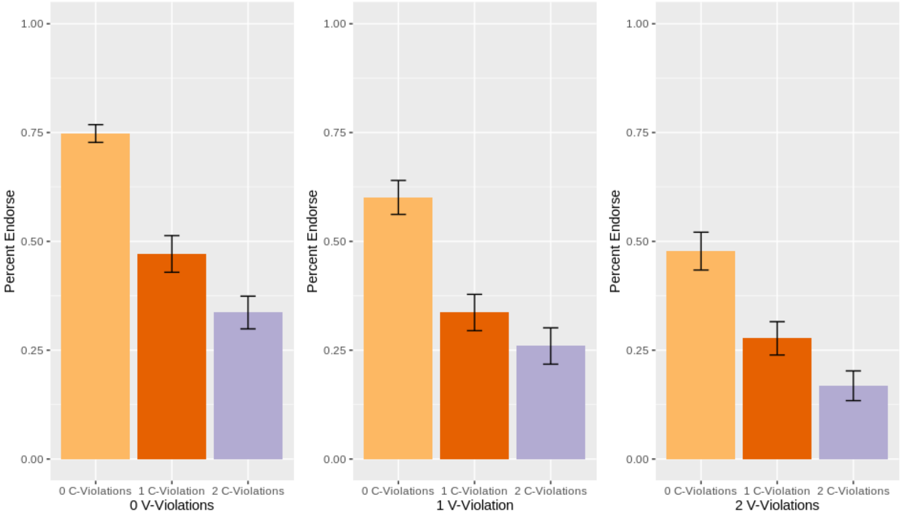

Figure 9: Results for the lexical decision task, Experiment 4. The vertical axis plots mean endorsement rate as a percentage, with standard error bars, and the horizontal axis divides the novel words according to their level of vowel-harmony violations, grouping by level of consonant violations. Note: *C-Violations* indicate violations of nasal harmony, and *V-Violations* indicate violations of backness harmony.

Figure 9 shows the results of the lexical decision task in Experiment 4. The logistic regression models contained a random intercept for subject and word. Two versions of this model were fit, the *alternative* model, with a three-level factor corresponding to violation level (0, 1, 2) of each phonotactic, and a *null* model, with a binary factor (*violating* vs. *non-violating*), discussed above. The alternative model fit the data significantly better than the null model ($\chi^2 = 19.928$, df(5), $p = 0.001$, assessed using the *anova()* function in R).

*8.3.2. Ratings task*

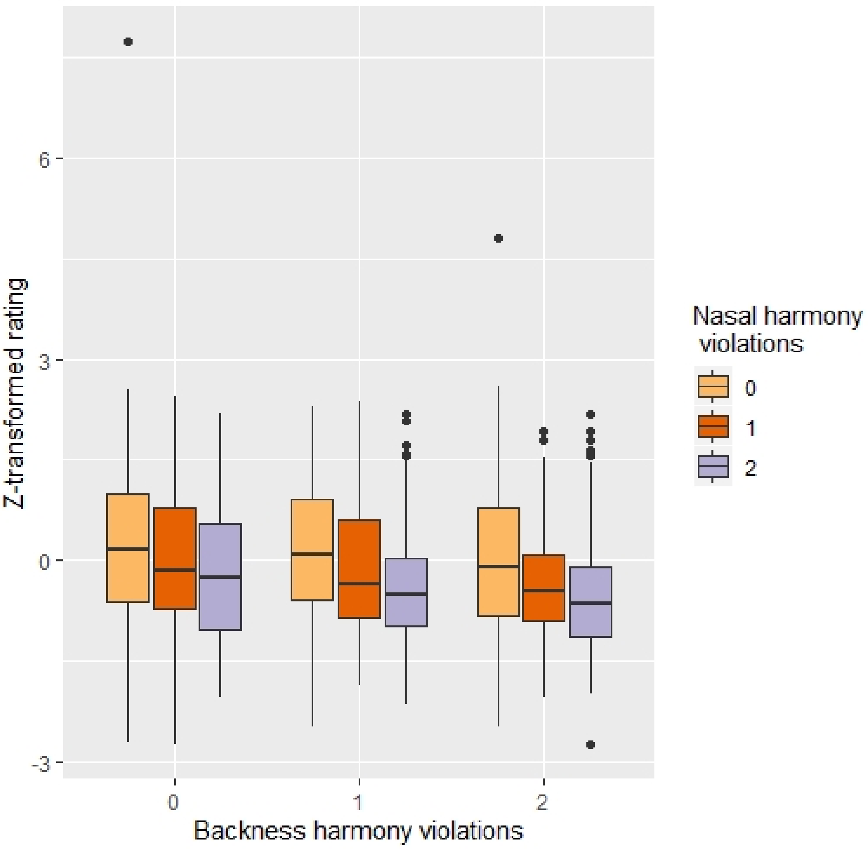

Figure 10: Results for the ratings task, Experiment 4.

Figure 10 shows the results of the ratings task in Experiment 4. A null and alternative model of the same structure as those described above were fit to the ratings data, with the alternative model yielding a significantly better fit ($\chi^{2} = 16.461$, df(5), $p = 0.006$).

*8.4. Local discussion*

Experiment 4 found that learners reliably distinguish between multiple levels of well-formedness (counting cumulativity, as in *pitetipe* (0 violations of nasal harmony) vs. *mitetipe* (1 violation) vs. *mitenipe* (2 violations)). These findings are in line with predictions of grammars which are capable of expressing cumulative relationships between constraint violations, and against predictions made by strict-ranking theories such as classical Optimality Theory.

## 9. Experimental discussion: Phonetic substance as a moderator of phonological learning

An unexpected yet repeated experimental finding was that participants exhibited an asymmetry in the strength of markedness associated with violating different phonotactics. In Experiments 1, 3a, 3b, and 4, violations of backness harmony were judged to be less severe than violations of nasal harmony. However, when backness harmony was paired with sibilant harmony in Experiment 2, no such disparity was observed. Why should this be?

One possible explanation is that fewer participants learned the backness harmony constraint than learned the nasal harmony constraint. I judge this hypothesis highly unlikely, because all participants had to pass an equal threshold for learning both phonotactics in the verification phase.

Another possibility is that the penalty for violating a phonotactic constraint is influenced by how perceptually distinct that constraint's conforming and non-conforming instantiations are. For example, 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.

This perceptual hypothesis could also explain 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. Possibly, learning two perceptually-subtle phonotactics was more difficult than learning one perceptually-subtle and one perceptually-salient phonotactics, leading to an overall reduction in accuracy in Experiment 2. The best way to capture this notion of "perceptual distinctiveness" — whether by counting distinctive features or by a more directly perceptual measure — is beyond the scope of this paper. However, the proposed phonetic explanation is in line with phonological theories which hold that both input statistics and perceptual similarity play a role in grammar (Steriade, 2001; Wilson, 2006; Zuraw, 2013; White, 2017, *i. a.*). Further research is necessary to disentangle the role of perceptual distinctiveness in modulating the severity of penalty associated with violation from possible increases in learning difficulty stemming from learning multiple perceptually-subtle phonotactics at once. Such research could also engage with formal measures of perceptual subtlety such as *confusion matrices*, though translating from such representations into grammar-internal constructs is a non-trivial task (for one approach, see White (2017)).

## 10. Modeling the experimental data: setup and training

In this section, I use the lexical decision task results from Experiments 1 and 4 to evaluate several phonological models: the Recursive Constraint Demotion (RCD) algorithm of Tesar and Smolensky (2000), a Stochastic OT model (Boersma and Hayes, 2001), and a Maximum Entropy model (Smolensky, 1986; Goldwater and Johnson, 2003).[11] Models were evaluated by finding the optimal settings to capture the results of Experiment 1, and then using these settings to predict endorsement rates for the stimuli in Experiment 4. To preview the results, while RCD was unable to capture any cumulativity, the Stochastic OT and MaxEnt models were able to qualitatively capture the ganging cumulativity in the generalization data from Experiment 1. However, the MaxEnt model generalized to the ganging- and counting-cumulativity seen in Experiment 4 substantially better than Stochastic OT.

[11] I chose to model the lexical decision task results because the phonological models considered here yield predictions about the likelihood of endorsement in a lexical decision task. To model the results of the ratings task, the models would require an additional linking hypothesis which treated numerical ratings as a probability of acceptance on a lexical decision task. While not unreasonable (cf. Breiss & Albright (*submitted*)), I address lexical decision data so as to minimize unnecessary assumptions.

### *10.1. The data to be fit: Experiment 1 revisited*

The results of the lexical decision task of Experiment 1 are reproduced below in Table 2. The center column lists the raw percentage of endorsement, while the right column provides the *scaled percent endorse*. Scaled percent endorse was obtained by summing the raw percent endorse over all categories of stimulus, and dividing each category's rate by that sum. This rescaling is necessary because the models pursued here yield predictions which differ from the format of the results in Experiment 1: they assign each to yield a probability distribution over four outcomes which sum to 1, rather than four probabilities of endorsement which each range between 0 and 1.

| *Violation profile (example)* | *Exp. 1 - unscaled* | *Exp. 1 - scaled* |
|---|---|---|
| Fully-conforming (*mimi*) | 75% | 39.0% |
| Backness-violating (*mimu*) | 58% | 30.2% |
| Nasal-violating (*mipi*) | 37% | 19.2% |
| Double-violating (*mipu*) | 22% | 11.4% |

Table 2: Generalization data from Experiment 1's lexical decision task, scaled and unscaled.

For modeling purposes, I adopt the simplifying assumption that the constraints FtBin (McCarthy and Prince, 1986; Prince, 1980), Max, Dep, and Id-[place] are undominated in all models throughout the learning process. Table 3 lists the exhaustive set of form types which was used in modeling; for example, the form [mimi] stands in for the experimental items such as [mimi], [mine], [nemi], [nene], etc..

### *10.2. Recursive Constraint Demotion*

I first evaluated the Recursive Constraint Demotion algorithm (RCD; Tesar and Smolensky (2000)).[12] This algorithm takes as input a series of *mark-data pairs* — a UR and a pair of candidates with a constraint that distinguishes them — and returns (if one exists) a constraint hierarchy that allows the optimal candidate to win. RCD constructs the ranking hierarchy by cancelling common violations among marked-data pairs, and then

[12] Variants on this approach, such as Biased Recursive Constraint Demotion (Prince and Tesar, 2004), and Low Faithfulness Constraint Demotion (Hayes, 2004) were also evaluated, and yielded the same result as classical RCD for the same analytic reasons.

| Phase presented | Stimulus profile | Stimulus |
|---|---|---|
| Training | *Fully-conforming* | [mimi] |
| ... | *Fully-conforming* | [pipi] |
| | *Fully-conforming* | [mumu] |
| | *Fully-conforming* | [pupu] |
| Generalization | *Backness-violating* | [mimu] |
| ... | *Backness-violating* | [pipu] |
| | *Backness-violating* | [mumi] |
| | *Backness-violating* | [pupi] |
| | *Nasal-violating* | [mipi] |
| | *Nasal-violating* | [pimi] |
| | *Nasal-violating* | [mupu] |
| | *Nasal-violating* | [pumu] |
| | *Doubly-violating* | [mipu] |
| | *Doubly-violating* | [pimu] |
| | *Doubly-violating* | [mupi] |
| | *Doubly-violating* | [pumi] |

Table 3: Simplified data used in modeling Experiment 1.

installing all constraints which prefer only winning members of the mark-data pair in the highest stratum. Then it removes those mark-data pairs for which the winning item has no violations un-cancelled. Then the algorithm repeats the ranking step, installing all unranked winner-preferring constraints in the second-highest stratum, removing all mark-data pairs with non-violating winners, and so on until there are no constraints left unranked.

The algorithm was provided with *mark-data pairs* consistent with the 32 fully-conforming stimuli from Experiment 1.[13] Constraints ranked were Agree([back]), Agree([nasal]), Id-[back], and Id-[nasal]; learning simulations were carried out in OTSoft (Hayes et al., 2017). The algorithm returned the following pairwise rankings: Agree([back]) >> Id([back]), Agree([nasal]) >> Id-[nasal], which implies the stratification below:

Stratum 1: Agree([back]), Agree([nasal])

[13] Here, the UR was always the well-formed form, and competitor SRs were created by altering the UR by one or two segments, so as to resemble items from the generalization phase.

Stratum 2: Id-[back], Id-[nasal]

As designed, the RCD does not yield a ranking which admits cumulativity. It does not allow a distinction among phonotactically-imperfect forms; that is, singly-violating *mimu* is not distinguished from doubly-violating *mipu*. Therefore I discard this framework as a suitable model of the phonological grammar as it is at work in the current experiments, and do not consider it when testing generalization to data from Experiment 4.

### *10.3. The Stochastic OT model*

Stochastic OT (Boersma et al., 1997) has been successfully employed to model a range of variable phonological phenomena (for an overview see Boersma (2003)), and is compatible with a range of phonological learning algorithms (Boersma and Hayes, 2001; Jäger, 2007; Magri, 2014; Magri and Storme, 2018, among others). A Stochastic OT grammar is learned incrementally, using an error-driven update rule to adjust ranking values for each constraint from their starting values until they minimize errors on the learning data. These ranking values then serve as the mean of a Gaussian distribution from which individual rankings are sampled when the grammar evaluates novel forms. For more details on the time course of learning/acquisition using the Gradual Learning Algorithm, an update rule used in Stochastic OT and the one employed in this model, see, for example, Escudero and Boersma (2003); Pater et al. (2007).

#### *10.3.1. Setting up the Stochastic OT model*

Stochastic OT is traditionally paired with a Markedness >> Faithfulness approach to phonotactics, in which the phonological grammar acts as a filter excluding all possible underlying forms except those which are consistent with the observed phonotactic distributions (Prince and Smolensky, 1993). However, this approach proves inappropriate for modeling the current experiment: when provided with categorical learning data of the form /mimi/ → [mimi], the Markedness >> Faithfulness setup predicts that participants will learn that *only* fully-conforming words are allowed (as the RCD model did) — the model does too well, and the initial ranking values remain unchanged. This is because the phonotactics-as-filter model never sees any informative learning data, specifically cases where faithfulness must be *promoted* to accommodate the attested forms.

This problem can be remedied, however, by appealing to the traditional notion of the Rich Base (Prince and Smolensky, 1993), with the role of the learning algorithm being to promote the markedness constraints which are responsible for the gaps in the learning data. Thus, the default ranking condition is Faithfulness >> Markedness, and the informative data are cases where faithfulness to the (disharmonic) UR is violated to produce the (harmonic) SR.

In principle, the full input to such a model would be all possible CVCV forms with C $\in$ {/p, m/} and V $\in$ {/i, u/} for each of the 32 training forms. To render each UR → SR mapping deterministic, I invoke the notion of initial-syllable faithfulness with the undominated constraint $\textsc{Id}_{\sigma 1}$ (cf. Becker et al. (2011)), along with the assumptions in section 10.1. Thus, for each of the resulting 16 tableaux, the winner was the fully-conforming candidate (ex., for the UR /pimu/, the winner was the fully-conforming candidate [pipi], beating out [pimu], [pimi], and [pipu]).

In this setup, the time-course of learning falls into three stages. Initially, while Markedness constraints are low and have not begun to approach the Faithfulness constraints on the ranking-value continuum, the algorithm mistakenly allows disharmonic URs such as /pimi/ to be realize without repair as [pimi]. However, if the algorithm is allowed to run its course, it will promote Markedness high enough so that there is no overlap in Gaussian distributions, and the model will only admit harmonic URs to be realized faithfully, and repair all others. Between these extrema is a zone where the ranking values of the Markedness constraints are close enough to those of the Faithfulness constraints that their associated Gaussian distributions overlap, and the algorithm sometimes allows disharmonic URs to be realized faithfully, and sometimes repairs them pursuant to the demands of the Markedness constraints: /pimi/ → [pimi $\sim$ pipi]. It is in this zone where non-categorical outcomes are permitted that the possibility for modeling cumulative constraint interaction lies.[14]

[14] A reviewer notes that the use of conjoined constraints (Smolensky, 1995) could also allow the model to capture cumulative results more easily. This is true, but in this case the failure would be one of explanation, rather than capacity — without justification for why learners should infer conjoined constraints in the first place, the addition of these constraints to the model is unmotivated.

### *10.3.2. Fitting the Stochastic OT model*

Rather than allowing the model to learn from the data, which would lead to an inversion of the initial ranking values and an ultimate return to the categorical Markedness >> Faithfulness regime as mentioned directly above, I tested the model at a range of fixed ranking values which fell within the intermediary zone of overlapping Gaussians discussed above. In order to capture the fact that experimental participants penalized violations of nasal harmony more than violations of backness harmony, I manipulated the difference in ranking values between the constraints which, if violated less often, would result in fewer attested violations of nasal harmony compared to backness harmony, and thus to a distribution over forms which put less probability mass on words which violate nasal harmony. Ranking values were fit by hand-adjusting the constraint weights down to a tenth of a point of ranking value to minimize the sum of the squared error between predicted and observed data. Note that while with the appropriate initial ranking values these values are *passed through* by the learner, I isolated them by hand here to yield greater precision than the step-size of the learning rule allowed. The ranking values which best fit the generalization data from Experiment 1 were Agree([nasal]) = 49.5, Agree([back]) = 48.5, Id-[nasal] = 50.5, and Id-[back] = 51.5. The predictions of the best-fitting model for the Experiment 1 data is presented in Table 4.

| Violations | Experiment 1 (scaled) | Stochastic OT |
|---:|---|---|
| None | 0.39 | 0.39 |
| Backness | 0.30 | 0.29 |
| Nasal | 0.19 | 0.18 |
| Backness & nasal | 0.11 | 0.14 |

Table 4: Stochastic OT model numerical fit to scaled Experiment 1 data.

Unlike the RCD model, Stochastic OT is able to qualitatively distinguish four rates of endorsement, and thus capture ganging cumulativity.

### *10.4. The Maximum Entropy (MaxEnt) model*

Like Stochastic OT, MaxEnt has been successfully used to model variable phonological phenomena. Unlike Stochastic OT, however, the MaxEnt model needs no modification to capture the experimental data. MaxEnt is most often used in combination with the “markedness only” theory of phonotactics, which takes the job of the learner as being to weight markedness constraints so as to maximize the likelihood of the attested distributions of data, rather than of a set of UR-to-SR mappings (cf. Hayes and Wilson (2008); Wilson and Gallagher (2018)). Adopting this theory allows for a straightforward implementation of the training phase in Experiment 1, in which participants were (implicitly) tasked with learning a grammar which maximized the likelihood of the fully-conforming data with which they were presented.

*10.4.1. Setting up the MaxEnt model*

I forced the model to match the non-categorical experimental results of the participants by placing a Gaussian prior with mean $(\mu) = 0$ and narrow standard deviation $(\sigma)$ on the weights of $\textsc{Agree}([\textsc{back}])$ and $\textsc{Agree}([\textsc{nasal}])$. This encourages the model to prefer solutions which have lower constraint weights in general, and thus causing it to depart from matching the categorical training data and towards predicting a distribution over multiple forms. Previous work using the MaxEnt framework has employed such priors to mimic the effect of biased learning in experimental outcomes (cf. White (2017); Wilson (2006)); here I use it to model the natural uncertainty about the final grammar that results from limited learning data.[15] To accommodate the asymmetry in severity between backness harmony and nasal harmony violations, I used differing values of $\sigma$ for the different markedness constraints: a larger standard deviation imposes less penalty for higher constraint weights and thus allows the model to give stronger penalties for violation of one markedness constraint compared to another.

*10.4.2. Fitting the MaxEnt model*

The candidate set consisted of a simplified tableau with the four types of stimulus used in training (fully-conforming, backness harmony-violating, nasal harmony-violating, and doubly-violating), with the 32 fully-conforming observations from the training phase. Values of $\sigma$ which best fit the experimental data were $\sigma_{\textsc{Agree}([\textsc{nasal}])} = 0.22$, and $\sigma_{\textsc{Agree}([\textsc{back}])} = 0.15$ (obtained by hand-adjustment to minimize the sum of the squared error between predicted and observed data); model predictions for the Experiment 1 generalization phase data are shown in Table 5 below.

[15] If this model were provided more training data it would overcome this uncertainty and learn a categorical grammar — it is hypothesized that the same would be true of the experimental participants, given sufficient time.

| Violations | Experiment 1 (scaled) | MaxEnt |
|---|---|---|
| None | 0.39 | 0.40 |
| Backness | 0.30 | 0.30 |
| Nasal | 0.19 | 0.17 |
| Backness & nasal | 0.11 | 0.13 |

Table 5: MaxEnt model numerical fit to scaled Experiment 1 data.

Model fit was qualitatively satisfactory, with the MaxEnt model predicting four distinct acceptance levels for the four categories of generalization stimuli.

## 11. Generalization to tetrasyllables

To this point, I have demonstrated that Stochastic OT and MaxEnt can qualitatively capture the ganging cumulativity observed in the generalization phase of Experiment 1, while the RCD model cannot. I now turn to evaluate Stochastic OT and MaxEnt on their ability to match human generalization on the data from Experiment 4, using the optimal parameters set from Experiment 1.

### *11.1. The data to be modeled: Experiment 4 revisited*

The unscaled generalization results from Experiment 4 are reproduced below in the leftmost column of Table 6. I evaluate the models on their ability to distinguish the nine violation profiles delimited by the three levels of violation of the two Agree constraints in the generalization phase of Experiment 4. I adopt a pairwise assessment of Agree violations (cf. Bakovic (2000); Pulleyblank (2002); Lombardi (1999); Tesar (2007); Kawahara (2011a)), and abstract away from the distinction between local and non-local violations (this issue is taken up again towards the end of the next section). For example, the novel word *mitepuni* incurred 3 violations of Agree[back] ([i. . . . . . u, e. . . u, u. . . i]), and 4 violations of Agree[nasal] ([m. . . t, m. . . . . . p, p. . . n, t. . . . . . n]).[16] This allows the modeling to remain in line with the statistical analysis in section 8.

[16] Although there is robust evidence that locality-based distinctions are attested in phonological *alternations*, for example in many harmony systems (cf. Suzuki (1998); Kimper (2011)), and have also been demonstrated in AGL experiments (cf. Finley (2017) for an overview), the question has yet to be addressed experimentally in the domain of phonotactics; I leave this as a question for future research.

### *11.2. Evaluating the generalization of all models*

I generated the predicted endorsement rates for Experiment 4's generalization stimuli using the settings obtained in section 10. To allow for a direct comparison between the mean percent endorsement in the experimental data and the predicted percent endorsement given by the models, I scaled the results of Experiment 4 to sum to 1. Table 6 shows the observed percent endorse for the violation profiles beside the predicted percent endorse from each of the models; these rates are plotted in Figure 11.

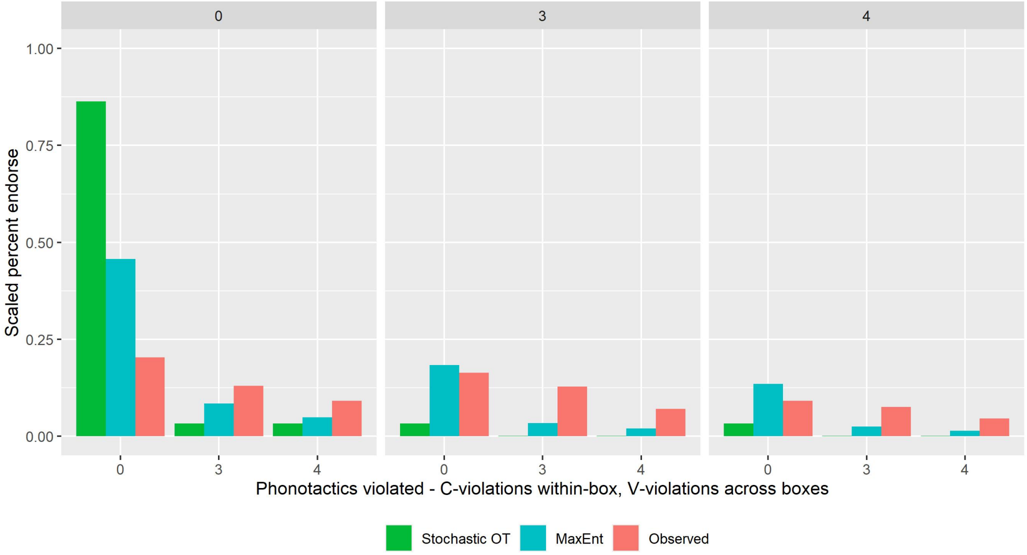

Figure 11: Stochastic OT (green) and MaxEnt (blue) model predictions for Experiment 4 lexical decision data (red). From left to right, the boxes are labeled with the Agree-[back] violation number, and columns from left to right within each box are labeled with the number of Agree-[nasal] violations at each level of Agree-[back] violation.

Model fits were evaluated by calculating the *sum of squared errors* for each of the violation profiles in Experiment 4. Log-likelihood values were not calculated because I am concerned with how closely each model approximates a target distribution given categorical training data and fixed parameter values, rather than using those parameter values to fit the model to a specific set of experimental outcomes. Model fits are evaluated numerically in Table 7.

At a glance, we can see that the models vary in their ability to match human generalization to the novel data. While both models can capture gang-

| Violation profile | Exp. 4 (scaled) | Stochastic OT (scaled) | MaxEnt |
|---|---|---|---|
| 0C, 0V (*pitetipe*) | 0.20 | 0.86 | 0.45 |
| 0C, 3V (*putetipe*) | 0.16 | 0.03 | 0.18 |
| 0C, 4V (*putetipo*) | 0.13 | 0.03 | 0.08 |
| 3C, 0V (*mitetipe*) | 0.13 | 0.001 | 0.03 |
| 3C, 3V (*mutetipe*) | 0.09 | 0.03 | 0.13 |
| 3C, 4V (*mutetipo*) | 0.07 | 0.001 | 0.02 |
| 4C, 0V (*mitenipe*) | 0.09 | 0.03 | 0.04 |
| 4C, 3V (*mutenipe*) | 0.07 | 0.001 | 0.02 |
| 4C, 4V (*mutenipo*) | 0.05 | 0.001 | 0.01 |

Table 6: Stochastic OT, and MaxEnt model numerical predictions for Experiment 4 lexical decision data. In the leftmost column, C and V indicate the number of violations of the consonant- and vowel-harmony phonotactics respectively, as in the parenthesized example form.

| Model | Sum of squared error |
|---|---|
| Stochastic OT | 0.498 |
| MaxEnt | 0.086 |

Table 7: Model comparison metrics using distance-agnostic AGREE.

ing cumulativity qualitatively — they predict different endorsement levels for each of the four categories distinguished by the binary violation status of AGREE([BACK]) and AGREE([NASAL]) — the same is not true of those violation profiles distinguished only by counting cumulativity. Stochastic OT predicts only four distinct levels of endorsement percentage across the nine attested rates of percent endorse in the data. MaxEnt, on the other hand, predicts nine unique rates. This difference follows directly from the fact that MaxEnt uses weighted constraints while the grammars underpinning the Stochastic OT models use strict-domination ranking. In ranked-constraint frameworks, multiple violations of a given constraint are no worse than a single violation (excluding cases of markedness ties), and therefore the distinction between stimuli of the form *pitetime* and *pinetime* is invisible to that grammar. MaxEnt distinguishes these forms because the harmony of each candidate is informed by not only *whether* but also *how often* each constraint is violated. Turning to the formal evaluation metrics MaxEnt performs better than Stochastic OT. Although neither matches the experimental data *extremely* closely, as would be expected if the models were allowed to adjust their parameters to fit the data, the goal of this evaluation task was to see how the models generalized from two- to four-syllable words, using the optimal parameter values learned from the data in the training phase, mirroring the human participants in Experiment 4.

Although in section 11.1 I used a distance-agnostic version of AGREE in order to ensure that the categories of outcome used to evaluate the grammars was supported by the statistical analysis of the experimental data, for the sake of completeness I carried out a parallel evaluation of the three models using a tier-local version of AGREE. For example, under this scheme the stimulus *mitepuni* incurred 2 violations of AGREE[BACK]-LOCAL ([m. . . t, p. . . n]) and 2 violations of AGREE[NASAL]-LOCAL ([e. . . u, u. . . i]), rather than the 3 and 4 respectively it incurred using the distance-agnostic AGREE. The results are in line with those reported in Table reftable:evalstatstable, and are presented below in Table 8.

| Model | Sum of squared error |
|---|---|
| Stochastic OT | 0.040 |
| MaxEnt | 0.006 |

Table 8: Model comparison metrics using AGREE-LOCAL.

## 12. General discussion

This study used a series of AGL experiments to investigate how learners acquire multiple phonotactic generalizations simultaneously, and how these generalizations interact in the grammar. 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. Experiment 2 replicated these findings using a different combination of phonotactics, sibilant-harmony and backness harmony, and Experiments 3a and 3b again replicated Experiment 1 using a training paradigm designed to more closely mimic natural first language acquisition. Experiment 4 asked participants to generalize their knowledge to longer words, and demonstrated that participants infer counting cumulativity as well. These results are significant first because they demonstrate cumulative effects on phonotactic well-formedness that cannot be explained by lexical frequency asymmetries. Further, they demonstrate that in the absence of evidence for or against constraint interaction, learners behave as expected if they have grammars in which constraint cumulativity is the norm, and in ways which are explicitly predicted to not be possible by grammars incapable of expressing cumulative relationships.

In the modeling section, these results were assessed in light of three contemporary phonological theories. I first demonstrated that traditional approaches to phonotactic learning (e.g. unaugmented Classical OT, Recursive Constraint Demotion (RCD) and its variants) were unable to capture even the ganging cumulativity observed in Experiment 1, a bar which the later two models were able to clear. Turning to these models, I compared the ability of a Stochastic OT model and a MaxEnt model to match the generalization behavior of participants in Experiment 4. Fitting the models to Experiment 1, where participants were asked to generalize from two-syllable exposure data to two-syllable testing data, I evaluated the models' fit to the four-syllable generalization data from Experiment 4 with the same settings. The results indicated that while both models were able to capture the ganging cumulativity in Experiments 1 and 4, only the MaxEnt model was able to capture both the counting and ganging cumulativity participants exhibited in Experiment 4.

## 13. Conclusion

Taken together, the experimental data and the modeling thereof indicate that both counting and ganging constraint cumulativity are structurally core to the phonological grammar. I suggest, therefore, that in as far as the purpose of phonological frameworks is to embody salient properties of the grammars humans use to speak and learn, weighted-constraint frameworks such as MaxEnt are to be preferred to other model types.

## References

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

Albright, A. (2012). Additive markedness interactions in phonology.

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

Bakovic, E. (2000). *Harmony, dominance and control.* PhD thesis.

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

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

Becker, M., Ketrez, N., and Nevins, A. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in turkish laryngeal alternations. *Language*, 87(1):84–125.

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

Boersma, P. (2003). Stochastic optimality theory. In *LSA Meeting in Atlanta January*, volume 3, page 2003. Citeseer.

Boersma, P. et al. (1997). How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 21, pages 43–58. Amsterdam.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, 32(1):45–86.

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

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

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

Durvasula, K. and Liter, A. (2020). There is a simplicity bias when generalizing from ambiguous data. *Phonology*.

Ernestus, M. and Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language*, pages 5–38.

Escudero, P. and Boersma, P. (2003). Modelling the perceptual development of phonological contrasts with optimality theory and the gradual learning algorithm. *University of Pennsylvania Working Papers in Linguistics*, 8(1):7.

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

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

Finley, S. (2015). Learning nonadjacent dependencies in phonology: Transparent vowels in vowel harmony. *Language*, 91(1):48.

Finley, S. (2017). Locality and harmony: Perspectives from artificial grammar learning. *Language and Linguistics Compass*, 11(1):e12233.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 144(852):285–307.

Frisch, S. A., Large, N. R., and Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of memory and language*, 42(4):481–496.

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

Gallagher, G. (2014). An identity bias in phonotactics: Evidence from cochabamba quechua. *Laboratory Phonology*, 5(3):337–378.

Goldwater, S. and Johnson, M. (2003). Learning ot constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, volume 111120.

Guy, G. R. (1997). Violable is variable: Optimality theory and linguistic variation. *Language Variation and Change*, 9(3):333–347.

Hansson, G. Ó. (2010). *Consonant harmony: Long-distance interactions in phonology*, volume 145. Univ of California Press.

Hayes, B. (2004). Phonological acquisition in optimality theory: the early stages. *Constraints in phonological acquisition*, pages 158–203.

Hayes, B., Tesar, B., and Zuraw, K. (2017). Otsoft 2.1, software package.

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

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., and Hothorn, M. T. (2016). Package 'multcomp'. *Simultaneous inference in general parametric models. Project for Statistical Computing, Vienna, Austria.*

Itô, J. and Mester, R.-A. (1986). The phonology of voicing in japanese: Theoretical consequences for morphological accessibility. *Linguistic inquiry*, pages 49–73.

Jäger, G. (2007). Maximum entropy models and stochastic optimality theory. *Architectures, rules, and preferences: variations on themes by Joan W. Bresnan. Stanford: CSLI*, pages 467–479.

Jäger, G. and Rosenbach, A. (2006). The winner takes it all—almost: Cumulativity in grammatical variation. *Linguistics*, 44(5):937–971.

Jarosz, G. and Rysling, A. (2017). Sonority sequencing in polish: The combined roles of prior bias & experience. In *Proceedings of the annual meetings on phonology*, volume 4.

Kawahara, S. (2011a). Aspects of japanese loanword devoicing. *Journal of East Asian Linguistics*, 20(2):169–194.

Kawahara, S. (2011b). Japanese loanword devoicing revisited: A rating study. *Natural Language & Linguistic Theory*, 29(3):705–723.

Kawahara, S. (2012). Lyman's law is active in loanwords and nonce words: Evidence from naturalness judgment studies. *Lingua*, 122(11):1193–1206.

Kawahara, S. (2013). Testing japanese loanword devoicing: Addressing task effects.

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

Kim, S. (2019). Modeling self super-gang effects in maxent: A case study on japanese rendaku.

Kimper, W. A. (2011). *Competing triggers: Transparency and opacity in vowel harmony*. University of Massachusetts Amherst Amherst, MA.

Kumagai, G. (2017). Super-additivity of ocp-nasal effect on the applicability of rendaku.

Lai, R. (2015). Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry*, 46(3):425–451.

Legendre, G., Miyata, Y., and Smolensky, P. (1990). *Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations*. Citeseer.

Lombardi, L. (1999). Positional faithfulness and voicing assimilation in optimality theory. *Natural Language & Linguistic Theory*, 17(2):267–302.

Magri, G. (2014). Tools for the robust analysis of error-driven ranking algorithms and their implications for modelling the child's acquisition of phonotactics. *Journal of Logic and Computation*, 24(1):135–186.

Magri, G. and Storme, B. (2018). Calibration of constraint promotion does not help with learning variation in stochastic optimality theory. *Linguistic Inquiry*, pages 1–27.

Martin, A. (2011). Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language*, 87(4):751–770.

Martin, A. T. (2007). *The evolving lexicon*. PhD thesis, University of California, Los Angeles Los Angeles, CA.

McCarthy, J. and Prince, A. (1986). Prosodic morphology, ms. *University of Massachusetts and Brandeis University*.

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

Nishimura, K. (2003). *Lyman's Law in loanwords*. PhD thesis, MA thesis, Nagoya University.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive science*, 33(6):999–1035.

Pater, J., Jesney, K., and Tessier, A.-M. (2007). Phonological acquisition as weighted constraint interaction. *Proceedings of the 2nd GALANA*, pages 339–50.

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

Prince, A. (1980). A metrical theory for estonian quantity', linguistic inquiry11, 511-62. *Prince51111Linguistic Inquiry1980*.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in phonology*, page 3.

Prince, A. and Tesar, B. (2004). Learning phonotactic distributions. *Constraints in phonological acquisition*, pages 245–291.

Pulleyblank, D. (2002). Harmony drivers: No disagreement allowed. In *Annual Meeting of the Berkeley Linguistics Society*, volume 28, pages 249–267.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Shademan, S. (2007). *Grammar and analogy in phonotactic well-formedness judgments*. PhD thesis, University of California, Los Angeles.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science.

Smolensky, P. (1995). On the internal structure of the constraint component con of ug. handout of talk presented at ucla, april 7, 1995. roa-86, rutgers optimality archive.

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

Suzuki, K. (1998). *A typological investigation of dissimilation*. PhD thesis.

Tesar, B. (2007). A comparison of lexicographic and linear numeric optimization using violation difference ratios. *Ms. Rutgers University*.

Tesar, B. and Smolensky, P. (2000). *Learnability in optimality theory*. Mit Press.

Walker, R. (2011). *Vowel patterns in language*, volume 130. Cambridge University Press.

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

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982.

Wilson, C. and Gallagher, G. (2018). Accidental gaps and surface-based phonotactic learning: A case study of south bolivian quechua. *Linguistic Inquiry*, 49(3):610–623.

Zuraw, K. (2013). 1* map constraints.

Zuraw, K. and Hayes, B. (2017). Intersecting constraint families: an argument for harmonic grammar. *Language*, 93(3):497–548.