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 cumulative constraint violations in phonotactics in the absence of a confounding natural-language lexicon. I find that learners consistently infer linear *counting* and *ganging* cumulativity between a range of phonotactic violations.

Keywords: phonotactics, cumulative constraint interaction, gang effects, artificial grammar, counting cumulativity, ganging cumulativity

1. Introduction

.

[⋆] 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 www.cbreiss.com/SupplementaryFiles/Breiss_ConstraintCumulativityInPhonotactics_SM.zip. Full responsibility for all remaining errors rests with the author. This work was supported by NSF Graduate Research Fellowship DGE-1650604 to Canaan Breiss.

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 (Albright, 2009, 2012; Pizzo, 2015; Durvasula and Liter, 2020). This method is inspired by work in experimental syntax (Featherston, 2005, 2019) where syntactic violations are manipulated in a crossed experimental design to tease apart the independent contribution of each one, and gain insight into how multiple violations are combined in the grammar. I combine this independent manipulation of violations with an Artificial Grammar Learning (AGL) paradigm which imposes a "sandbox" environment on the learner. 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 learned grammar, rather than asymmetrical distributions of structures in the lexicon; this point is taken up again in section 3.3. Note that the use of an artificial language does not render the experimental results impervious to the influence of whatever non-linguistic cognitive factors may be at play in acceptability judgements. We expect such effects (though do not model them here explicitly), but do not anticipate such effects exerting an asymmetrical effect on different items in the experiment, so the within-experiment comparisons which are the focus of this paper should be unbiased.

To preview the results, I show that learners consistently infer linear *counting* and *ganging* cumulativity between a range of phonotactic violations. I discuss the compatibility of these results with a range of contemporary constraint-based phonological frameworks, and argue that only probabilistic weighted-constraint frameworks such as Maximum Entropy Harmonic Grammar (Smolensky, 1986; Goldwater and Johnson, 2003) and Noisy Harmonic Grammar (Boersma and Pater, 2008) are able to capture both counting and ganging cumulativity.

2. Constraint cumulativity in phonological theory

Constraint-based phonological frameworks diverge on whether they can model cumulative constraint interactions. Classic 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 a Classic OT tableau.

In HG the optimal outcome is the one which has the lowest harmony penalty 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, with "H" denoting Harmony.

Thus, HG and OT sometimes predict different outcomes from the same schematic example because candidates' violations are cumulative in HG but

[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.

not in OT. Although in OT constraint violations can be compared numerically when two candidates tie on all higher-ranked constraints, these cases are not considered cases of cumulative constraint interaction. On the other hand, HG *cannot help* but exhibit cumulativity, regardless of the relative strengths of the constraints involved.

Jäger and Rosenbach (2006) delineate two possible types of constraint cumulativity: *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

In this section I review cases of alternations and phonotactic distributions which are suggestive of cumulative constraint interaction, and in many cases have been used to support arguments in favor of such analyses. Data patterns that can be analyzed in terms of cumulative constraint interaction in frameworks that allow it, such as HG, have often been analyzed in OT by means of Local Constraint Conjunction (LCC) (Smolensky, 1993; Smolensky and Legendre, 2006), a formal mechanism that encodes specific instances of cumulative interaction in a ranked-constraint model by brute force.[2]

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; that is, alternations (Goldwater and Johnson, 2003; Coetzee and Pater, 2006; Pater, 2009; Zuraw and Hayes, 2017, *i. a.*). The majority of phonological

[2] Note, however, that LCC and HG are not necessarily incompatible (cf. Shih (2017)), but the role LCC plays in HG is quite different than the way it is used in OT, and therefore will not be discussed further here.

patterns which are suggestive of a cumulative analysis are *ganging cumulativity*, where two (or more) distinct factors interact to determine how the SR for a given UR is realized. 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 the static phonotactic restriction known as "Lyman's Law" which prohibits multiple voiced obstruents within a word can be construed as a case of constraint cumulativity.[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), and Kawahara (2013) follow 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 voicing alternation triggered by compound formation in the native Japanese lexicon known as *rendaku* (Kawahara, 2012) in a further case of apparent cumulativity, supporting the observations made by Itô and Mester (1986). Recent studies by Kawahara (*to appear*) and Kawahara and Breiss (*resubmitted*) 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) and Kumagai (2017) demonstrate that only two nasals (but not one) block *rendaku* application in Japanese compounds.

3.2. Constraint cumulativity in phonotactics

Data which suggest cumulative constraint interaction have also been noted in phonotactics, generally taking the form of additive effects of multiple

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

marked structures on the likelihood of lexical attestation or experimentally-assessed acceptability. Durvasula and Liter (2020) used an AGL paradigm to examine what kinds of generalizations were formed during phonotactic learning, and found that speakers learn multiple generalizations consistent with their data. Crucially, they also found that when asked to extend these generalizations to novel items, the generalizations interacted in a cumulative manner: a novel stimulus violating two phonotactics was more likely to be rejected than one violating only one.

Other studies on cumulative effects in phonotactics come from the study of natural languages. One such piece of evidence comes from Pizzo (2015). Pizzo carried out a series of large-scale acceptability judgement studies on the cumulative effects of syllable margin well-formedness constraints in English. She found that nonce words that violate English syllable-margin phonotactics once, ex. *plavb* or *tlag*, were judged less well formed than those which did not violate — *plag* — and crucially *more* well formed than those which violated twice, ex. *tlavb*. Albright (2009) used a different method, modeling the experimental acceptability of a range of nonwords containing a range of structures with differing well-formedness. He found that models which took into account multiple marked structures in a word were a better fit for two existing datasets of human judgements than those which took account of only one such structure per word, suggesting that the experimental participants judged the well-formedness of a nonce-word based on the cumulative well-formedness of its structures. Taken at face value, these studies constitute suggestive 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 violation alone.

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

While suggestive of cumulative behavior, however, the findings of Pizzo and Albright have an alternative explanation. This is because in these cases, experimentally-determined well-formedness is highly correlated with the frequency of such structures 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 prominent role of lexical statistics in influencing well-formedness judgments is well established. 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 more 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 causality — 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) and 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 lexical statistics can be carefully controlled, as was done by Durvasula and Liter (2020). All phonemes were equally frequent in the training data, as were all local phoneme bigrams and all non-local phoneme bigrams on C- or V-tiers. 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.

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)). 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 consonant nasality harmony (hereafter *nasal harmony*): in conforming words, all 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*: in conforming words, 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*: in conforming words, 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 categorical well-formedness judgements in what I term the *binary 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 binary choice and ratings tasks.[4]

[4] An anonymous reviewer raises the possibility that participants could simply be judging the well-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 the 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-phonological" 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 for the present paper (though see Durvasula and Liter (2020) for recent work focusing on exactly what level of representational granularity learners form

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 the 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*, *mumu*, 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 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 *po*$\underline{t}$*u* vs. *po*$\underline{n}$*u*.

In piloting, participants showed a strong preference for forms with identical consonants or vowels despite there not being any more stimuli containing a given identical pair of phonemes than any other combination of phonemes

generalizations over in AGL experiments). Interested readers can access the raw data from all experiments reported here in the supplementary materials.

(in line with the general findings of Gallagher (2013)). 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., *tot*$\underline{u}$ vs. *tot*$\underline{i}$). 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 nonwords 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 binary 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 binary decision task and the ratings task. Both tasks used the same set of novel words. In the binary 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 obtaining informed consent from the participants, 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 binary 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 binary decision task I modeled the log-odds of endorsing an item as a function of its phonotactic violations using mixed-effects logistic regression. Note that since this task involves a binary outcome, I probe the models for cumulativity on the log-odds scale, rather than the probability scale. For the ratings task I modeled the raw numerical data using mixed-effects linear regression.[5]

Regardless of the domain of analysis — log-odds of endorsement or numerical rating — once a model was fit, I explored the interaction term using planned comparisons 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). Probing this difference is important, since it is here where cumulativity (or lack thereof) can be established. As discussed in section 2, a lack of cumulativity in constraint interactions is characterized by a single violation of the highest-ranking constraint being prioritized

[5] What exact calculation subjects were performing to give their response to this question is a relevant question, but is beyond the scope of this paper. The experimental prompt, *"How good does X sound as a word in the language you've learned?"*, could have been interpreted either as a request for a similarity score (of unknown parameterization) or as a request for a probability of membership, and could also have differed among subjects. The choice of linear model here is informed by the fact that many subjects in the debriefing talked about giving words a greater or smaller number of "points," which suggests the use of a numeric scale, but the topic is open to further inqury.

to the exclusion of all others. The interaction term in the model is a way of measuring this, since it can indicate whether the effect of one violation differs depending on whether it is accompanied by another violation or not. We expect that if the effect of one violation is “cancelled out” in the presence of another, indicating a *lack* of cumulativity, this will be indicated by a significant interaction term with a positive coefficient. If, on the other hand, the main effects of phonotactic violation are significant and the interaction of the two is *not*, we cannot conclude that participants inferred anything but a decrement in log-odds of acceptance or in numerical rating for violating each constraint independently. If under this scenario the post-hoc tests reveal significant differences (in log-odds of endorsement or numerical rating) between each of the singly-violating levels and the doubly-violating level, this is robust support for cumulative constraint interaction at work. However, if the two main effects of phonotactic violation are significant, their interaction is not, and also the post-hoc comparisons between singly-violating levels and the doubly-violating level is *not* significant, we cannot conclude that there was *no* cumulativity inferred, but neither does the experiment provide strong supporting evidence.

Note that although I regressed on the raw ratings data, for the sake of legibility I plot z-transformed ratings throughout the paper. These ratings were obtained by subtracting the mean rating for each subject from all of the ratings for that subject, and then dividing the result by the standard deviation of the ratings for that subject.

5.3. Results

5.3.1. Binary decision task

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

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

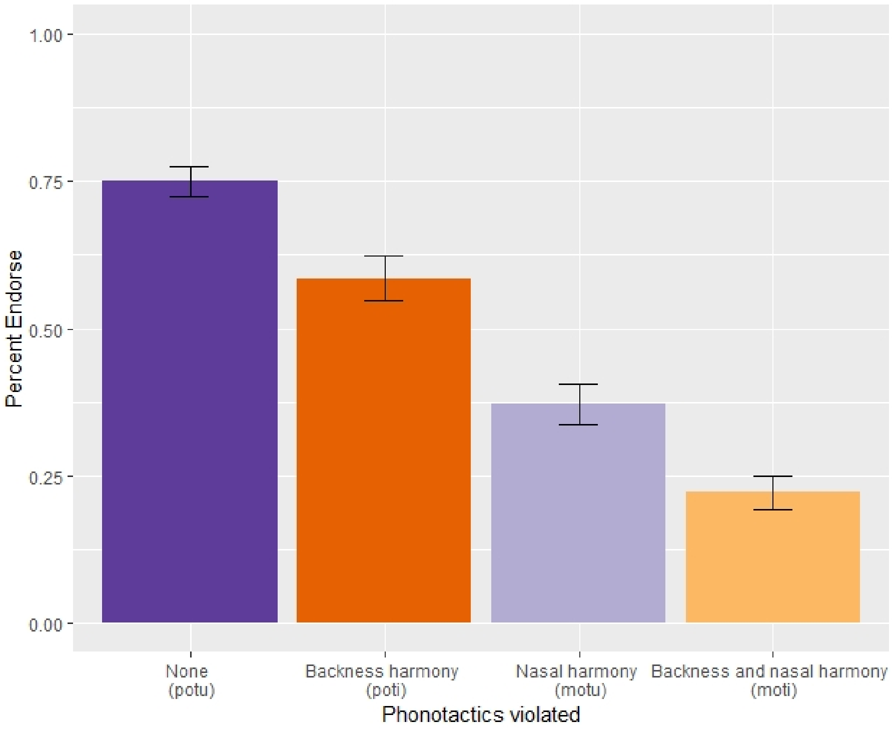

Figure 2: Results for the binary decision task, Experiment 1. The vertical axis plots mean endorsement rate — the likelihood of an individual item of a given profile being judged as being able to be a part of the language in question — as a percentage with standard error bars, and the horizontal axis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

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]

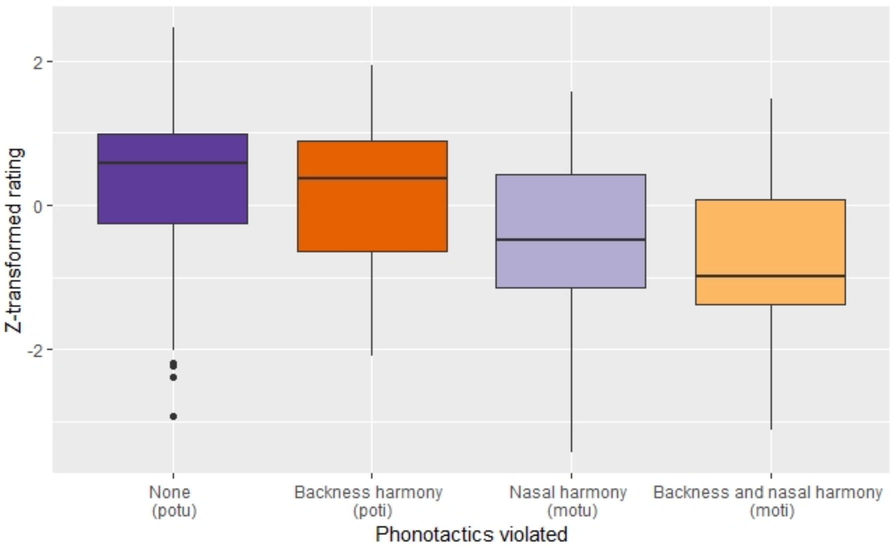

Figure 3: Results for the ratings task, Experiment 1. For the remainder of the paper, the central line in each boxplot indicates the median, with the colored portion of the box extending from the 25th percentile to the 75th; whiskers extend a further 1.5 times the inter-quartile range of the data. For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

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 binary 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

[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.

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 sandbox environment, learners infer ganging cumulativity between violations of two separate phonotactic constraints in their learning data. In the binary decision task, doubly-violating 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 without a significant interaction, but since only backness-violating forms differed from doubly-violating forms in the post-hoc tests, the evidence on this task is slightly weaker. 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 participant exclusion rate, about 50%. Although it is an open question as to why this specific experiment should have such a high exclusion rate — raising the possibility, though by no means the certainty, that a substantive difference between the sibilant harmony and nasal harmony could be the cause — this question deserves experimental inquiry beyond the scope of this paper. For the present purposes, though the exclusion rate is high, 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, and so we can trust the results of the experiment in as far as they are informative about cumulativity.

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 binary decision task contained only one presentation of each of the novel words, rather than two, to shorten the experiment and remove the between-block dependency noted in footnote 6.

6.2. Results

6.2.1. Binary decision task

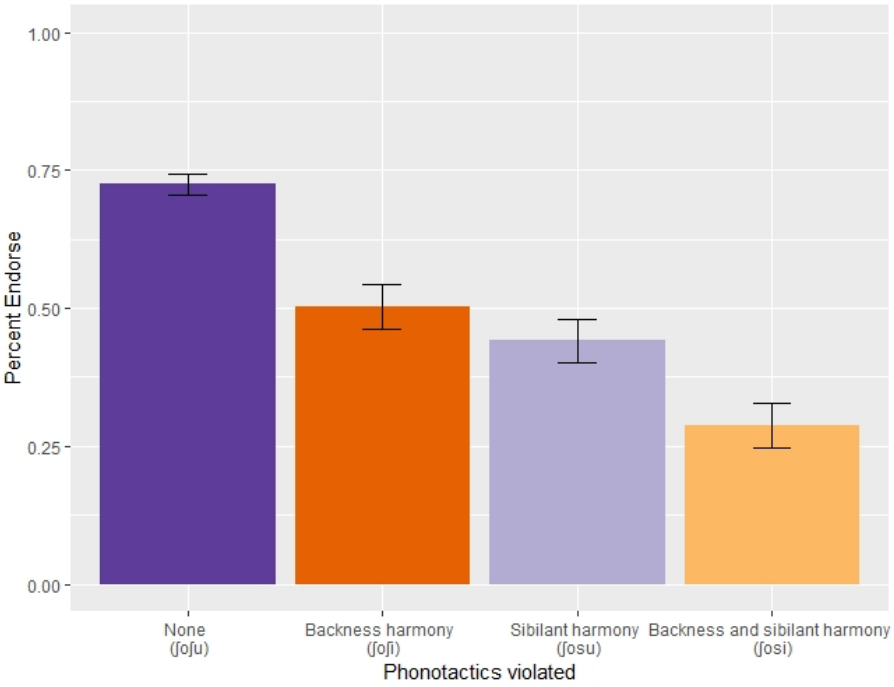

Figure 4: Results for the binary decision task, Experiment 2.

Figure 4 shows the results of the binary 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. $= 0.414$, $z = 3.018$, $p = 0.003$).

6.2.2. Ratings task

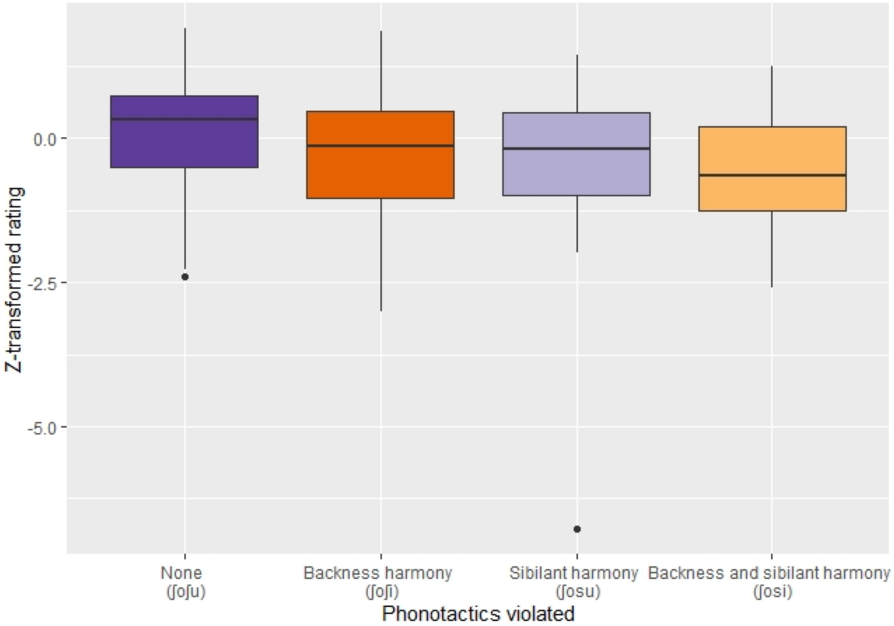

Figure 5: Results for the ratings task, Experiment 2.

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 binary 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 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 different types of phonotactic constraints. Although it is beyond the scope of this paper, examining a wider range of phonotactics and how they engage in cumulative behavior would be a valuable contribution to the empirical literature on cumulativity.

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 which contrasted in consonant harmony violation only, and vice versa. 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. The passive exposure training led 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 a failure to learn both phonotactics to criterion.

Because of the longer exposure phase, the binary 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 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. Binary decision task

Figure 6 shows the results of the binary 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 violating nasal harmony ($\beta = -1.562$, std. err. $= 2.227$, $z = -6.900$, $p < 0.001$); the interaction between the two was

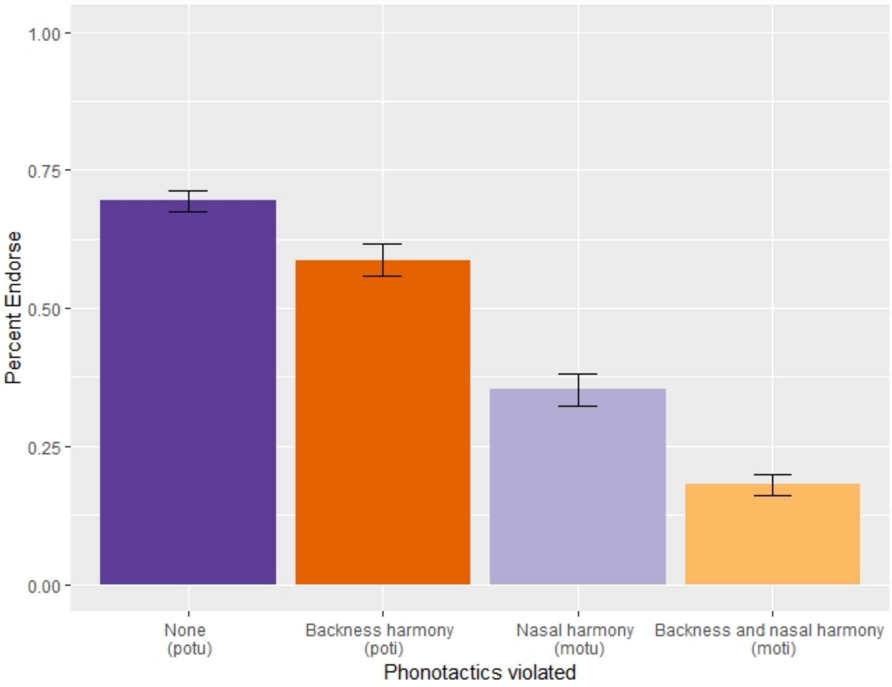

Figure 6: Results for the binary decision task, Experiment 3a.

not significant (β = -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 (β = 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 (β = 0.034, std. err. = 0.533, z = 0.065, p = 0.948).

7.2.2. Ratings task

Results of the ratings task are presented in Figure 7. The main effect of violating nasal harmony was significant (β = -17.536, std. err. = 3.707, z = -4.730, $p < 0.001$), but the main effect of violating backness harmony was not (β = 0.706, std. err. = 3.706, z = 0.190, p = 0.850), nor was the interaction between these factors (β = -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.

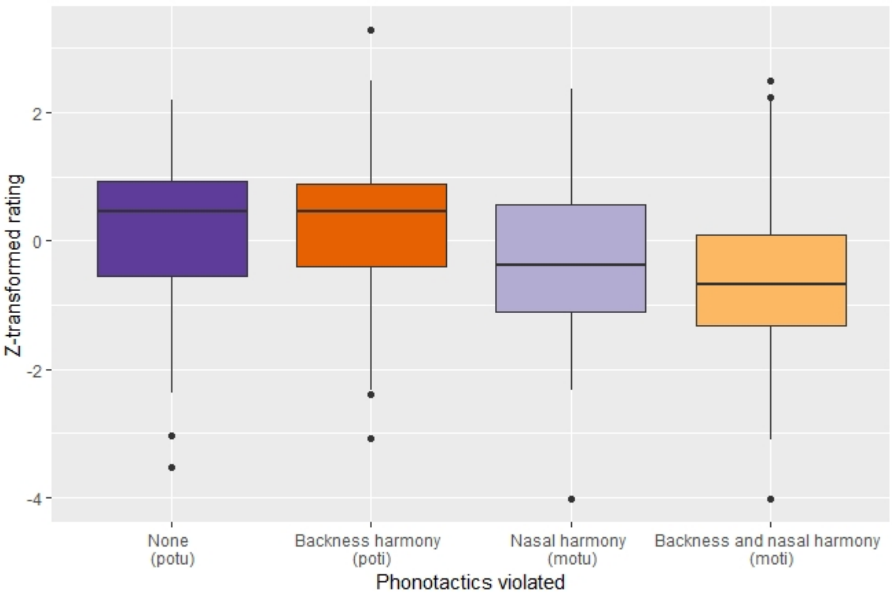

Figure 7: Results for the ratings task, Experiment 3a.

7.3. Local discussion

Experiment 3a provides some evidence for the robustness of inferred cumulativity under more naturalistic passive exposure training: in the binary 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

In an attempt to better understand 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 again observe 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 binary decision task during the test phase.

7.5. Results I: Do subjects infer ganging cumulativity?

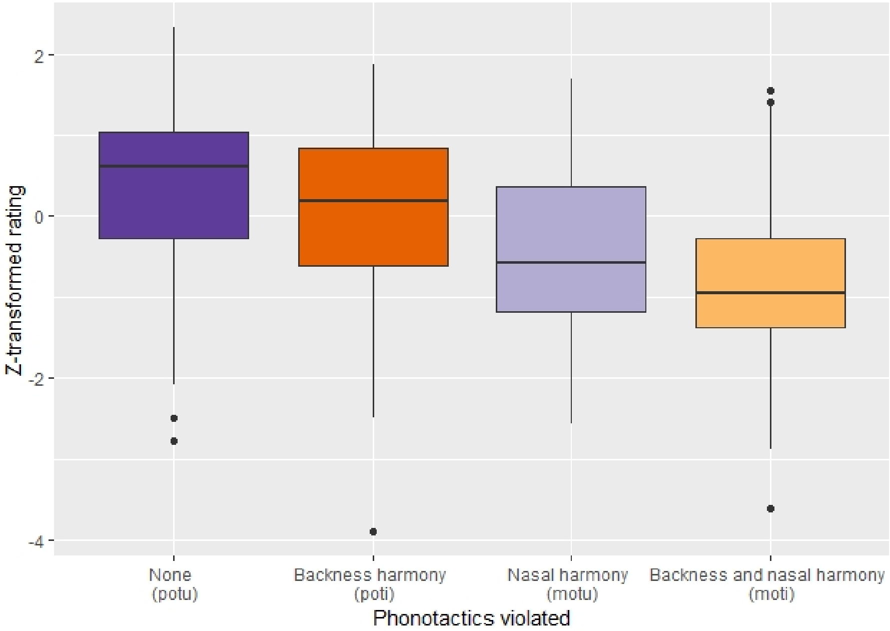

Figure 8: Results for the ratings task, Experiment 3b.

Results of the ratings task are presented in Figure 8. Mirroring results of the binary 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. $= 9.089$, $z = 1.149$, $p = 0.251$).

7.6. Local discussion

Experiment 3b fails to replicate the null effect observed in the ratings task in Experiment 3a - that is, as in Experiments 1 and 2, Experiment 3b showed evidence of ganging cumulativity. I take this to support the hypothesis laid out in section 7.3 that the lack of cumulativity observed in Experiment 3a was due to random experimental variation and not to a substantial difference in the experimental design.

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, it 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 taught the exact same two-syllable language used in Experiment 1, but 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 see 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[7]) of both phonotactics, for six stimuli per violation-level combination cell.

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.[8]

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.

[7] Note that in this paper I refer to "locus of violation", indicating the presence of a non-majority-matching segment quality, rather than violations of particular constraints. I leave for future investigation the possible utility of counting cumulativity in probing what exactly counts as a violation of agreement, for a variety of definitions of harmony-enforcing constraints.

[8] An anonymous 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.

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 *null model* that contained two binary fixed effects (whether or not a form violated each phonotactic (*yes/no*)), to an *alternative model* which contained two three-level factors denoting *how many* times a form violated each phonotactic (0, 1, 2), and compared these non-nested models using the Akaike Information Criterion (AIC) (Burnham and Anderson, 2002, 2004). The difference in AIC values between the alternative and null models can be converted into an odds ratio that the alternative model is the one with more explanatory power than the null model. This statistical test directly corresponds to the question that is at stake in linguistic theory: is a model which allows counting cumulativity more likely, given this data, than one which does not?

8.3. Results

8.3.1. Binary decision task

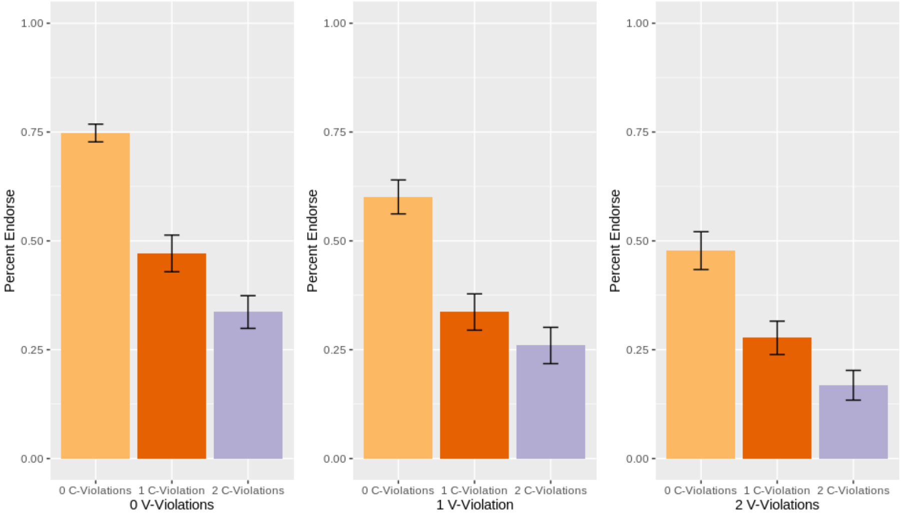

Figure 9: Results for the binary decision task, Experiment 4. The vertical axis plots mean endorsement rate as a percentage, with standard error bars, and the horizontal axis divides the novel words according to their level of vowel-harmony violations, grouping by level of consonant violations. Note: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

Figure 9 shows the results of the binary 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 only difference between these two models was that the alternative model allows for a distinction between multiple levels of violation — counting cumulativity — and the null does not. I found that the odds of the alternative model being superior are ≈ 141:1 (ΔAIC = 9.9). I take this as evidence in favor of counting cumulativity playing a role in generating the experimental data.

8.3.2. Ratings task

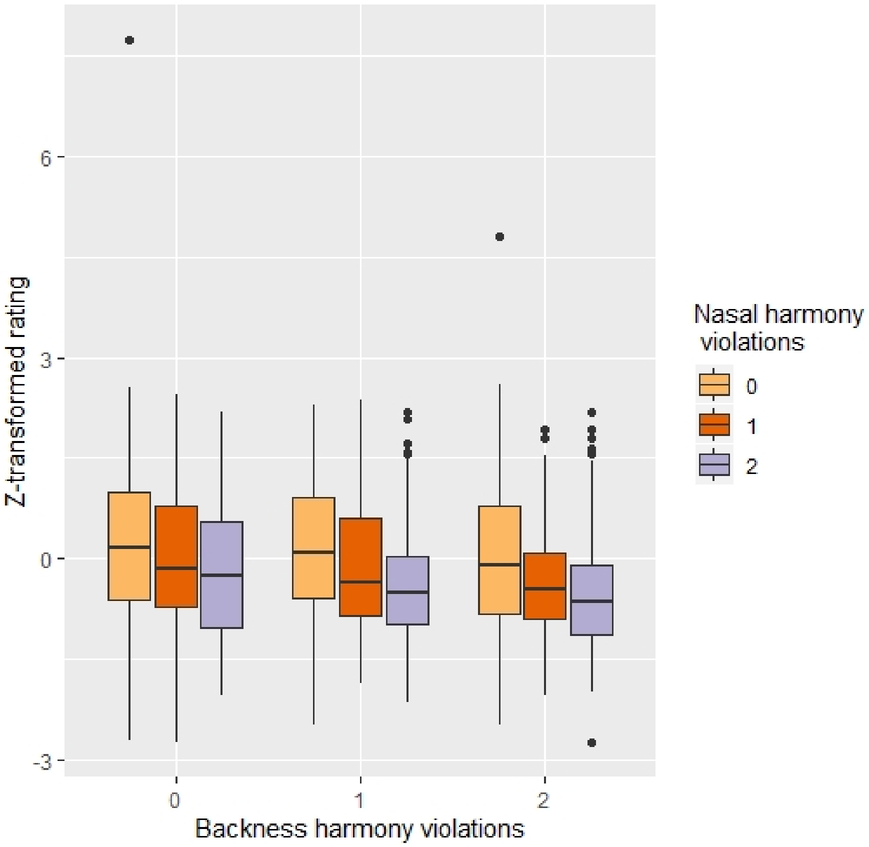

Figure 10: Results for the ratings task, Experiment 4.

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, and AIC-based model comparison indicated that the odds of the alternative model being superior were $\approx$ 33:1 (ΔAIC = 7), again supporting cumulativity.

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 Classic Optimality Theory.

9. Discussion and conclusion

This paper 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. The potential significance of these results is that 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.

9.1. Implications for phonological frameworks

As discussed in section 2, phonological frameworks differ in their generative capacity to capture cumulative constraint interactions. By design, Classic OT and other strict-domination-based frameworks rule out both counting and ganging cumulativity. On exception to this rule is Stochastic OT (Boersma et al., 1997; Boersma and Hayes, 2001), which, while able to capture ganging cumulativity within a certain range (cf. Zuraw and Hayes (2017); Smith and Pater (2017); Kawahara (*to appear*)), is unable to capture counting cumulativity for the same reason that Classic OT is unable to — a single constraint is either violated or satisfied, and no notion of "number of times violated" exists beyond what is needed to determine which constraint is violated *more* often in case no candidates are violation-free (Prince and Smolensky, 1993, p. 18). While in-depth model comparison is not carried out here, this data suggests the tentative conclusion that only weighted-constraint frameworks such as Maximum Entropy Harmonic Grammar (Smolensky, 1986; Goldwater and Johnson, 2003) and Noisy Harmonic Grammar (Boersma and Pater, 2008) are adequately expressive models of the phonological grammar.

9.2. Future work

In the long run, it would be sensible to evaluate these frameworks not just in coarse, qualitative terms, but in their ability directly to predict the results of experiments such as the ones described above. Such predictions, however, require more than just a set of competing phonological frameworks; we need explicit and well-supported linking hypotheses that relate the output of phonotactic grammars (fitted to the training data of the experiment) to the participant responses. The development and experimental validation of such mechanisms remains a topic for future research.

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.

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

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.

Boersma, P. and Pater, J. (2008). Convergence properties of a gradual learning algorithm for harmonic grammar.

Burnham, K. P. and Anderson, D. R. (2002). Model selection and.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

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 generalising from ambiguous data. *Phonology*, 37(2):177–213.

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.

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. (2013). Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, pages 253–295.

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.

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. 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.

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.

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.

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.

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

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

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.

Shih, S. S. (2017). Constraint conjunction in weighted probabilistic grammar. *Phonology*, 34(2):243–268.

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

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

Smolensky, P. (1993). Harmony, markedness, and phonological activity. In *Rutgers optimality workshop*, volume 1, pages 87–0000.

Smolensky, P. and Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT press.

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

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