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 that 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 — Multiple Grammars (based on the classical Recursive Constraint Demotion algorithm (Tesar and Smolensky, 2000)), Stochastic OT (Boersma and Hayes, 2001) and Maximum Entropy HG (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 all three frameworks are able to capture the ganging cumulativity participants displayed, the Maximum Entropy model best captures both counting and ganging cumulativity. This follows directly from MaxEnt's use of weighted, rather than ranked, constraints.

*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, Sara Finley, Shigeto Kawahara, James White, Colin Wilson, three anonymous reviewers, and members the audience at the LSA Annual Meeting 2019, SCAMP 2019, and the UCLA Phonology seminar for much helpful commentary and discussion. Thanks also to Beth Sturman and Eleanor Glewwe for recording stimuli; to Henry Tehrani for technical assistance; and to Lakenya Riley, Grace Hon, Amanda Singleton, Shreya Donepudi, Azadeh Safakish, and Ruby Dennis for helping to run subjects. Full responsibility for all remaining errors rests with the author. This work was supported by NSF Graduate Research Fellowship DGE-1650604 to Canaan Breiss.

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 a subject of ongoing debate in phonology. Cumulativity is often discussed in contexts where the grammar seems to reference continuous values along a scale, or else "counts" phonological structures. For example, in Kaska (Athabaskan; Lionnet (2016)), /ɛ/ lowers to [a] only when its syllable ends with [h] and immediately precedes a nuclear [a], but not when each of those conditions are met independently:

/sɛ=hta:n/ → [sa̲hta:n]	*"(s)he put (long object) there"*	*ɛh, *ɛC$_0$a
/sɛ=ta:n/ → [sɛ̲ta:n]	*"(long object) is there"*	*ɛC$_0$a
/sɛ−h=tsu:ts/ → [sɛ̲htsu:ts]	*"(s)he put (fabric) there"*	*ɛh

Cumulative effects are also attested in acceptability judgements: Pizzo (2015) 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*.

In this paper I take on the question 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 Featherston's 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 — Multiple Grammars (based on 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. 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 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. This is because the two violations of Constraint 2, though tolerated individually, together outweigh the penalty associated with the single violation of Constraint 1.

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

Figure 1: Schematic example of a HG tableau.

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

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 takes the form 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.*). For instance, the additive 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 cumulativity. Pater finds that while speakers tolerate voiced obstruents and geminate consonants unrepaired when adapting loan words, they preferentially repair words which contain voiced geminates by devoicing 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 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. Recent studies by Kawahara (*ms*) also indicate that the relationship between form and meaning characteristic of sound-symbolism displays cumulative effects.

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 how frequent these 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 prominent role of lexical statistics in influencing well-formedness judgments is well established in generative phonology (Frisch et al. (2000); Shademan (2007); Daland et al. (2011); Jarosz and Rysling (2017), among many others). Pioneering work by Coleman and Pierrehumbert (1997) highlighted the connections between the lexicon and phonotactic well-formedness in their predictive model of nonword judgments. 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.

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, 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. 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 consonant sibilancy harmony (*sibilant harmony*): 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 harmony (*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 completed a *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 should be the result of converging evidence from both dependent variables.

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

In piloting, participants showed a strong preference for forms with identical consonants or vowels despite no numerical advantage for these forms in the training data. Therefore, the verification trials were balanced so that pairs whose fully-conforming word had identical consonants (ex. *totu*) differed only in their violation of backness harmony (ex., *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 lexical decision task and then the ratings task. The exposure phase consisted of two blocks of 32 pseudo-randomized self-paced trials. During the exposure phase, similar-sounding items were presented together in blocks of eight, with each subject assigned at random to one of four counter-balancing orders of the four blocks. For example, in one counterbalancing group, participants first heard eight words with front vowels and voiceless stops (ex. *peti, tipi, tepe, piti...*) followed by eight words with back vowels and nasal stops (*monu, nunu, mumo, numo...*), followed by eight words with back vowels and voiceless stops (*topu, pupo, topo, putu...*) followed by eight words with front vowels and nasal stops (*nini, meni, nemi, mene...*).

After the exposure phase, participants completed 16 self-paced two-alternative forced choice verification trials, 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

The experiment was conducted 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 (**?**) 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 specific to the analysis, and random slopes for all fixed effects by subject. In cases of non-convergence, interactions among random slopes were removed first, then the slopes themselves, until the model converged. For the lexical decision task I modeled the log-odds of endorsing an item using logistic regression, and for the ratings task I modeled the raw numerical data using a linear model. Note that although I regressed on the raw ratings data, for the sake of legibility I plot z-normalized ratings throughout the paper.

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

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

5.3.1. Lexical decision task

Figure 2 shows the results of the lexical decision task in Experiment 1. The final logistic regression model included a random intercept for subject and word, and a four-level fixed effect for violation profile. The log-odds of endorsement differed significantly between the fully-conforming forms and the nasal harmony-violating forms ($\beta = -1.748$, $p < 0.001$), and the backness harmony-violating forms ($\beta = -0.813$, $p < 0.001$). Participants endorsed doubly-violating forms with a significantly lower likelihood than nasal harmony-violating forms ($\beta = -0.793$, $p = 0.002$) and backness harmony-violating forms ($\beta = -1.728$, $p < 0.001$).

5.3.2. Ratings task

Results of the ratings task are presented in Figure 3. The final model included a four-level factor of violation profile, and a random intercept for subject and word. Forms violating the backness harmony phonotactic were rated significantly lower than fully-conforming forms ($\beta = -7.100$, $p = 0.019$) and forms violating the nasal harmony phonotactic were rated significantly

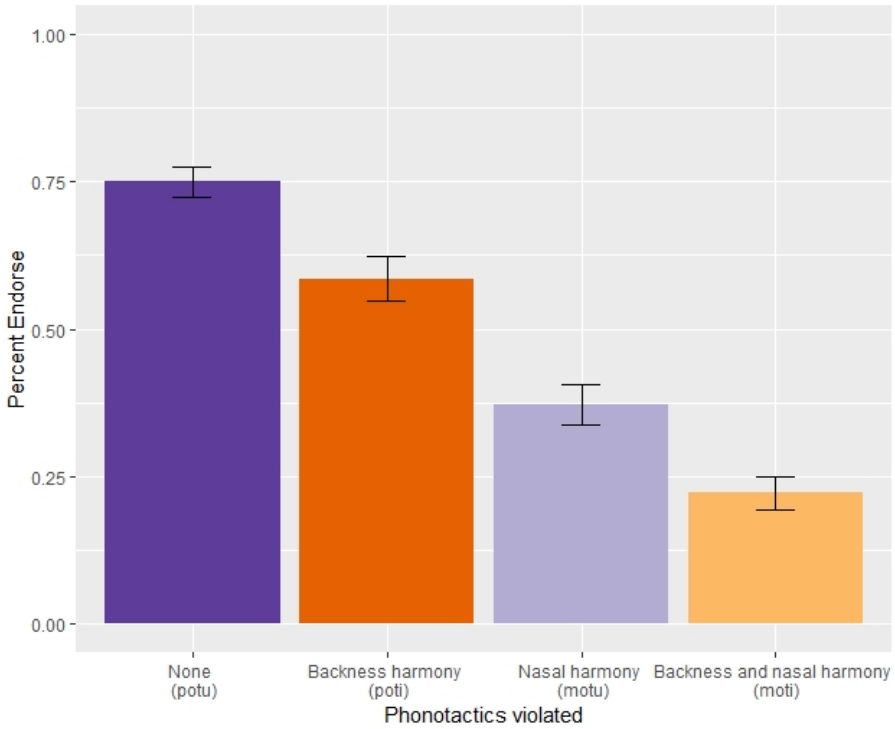

Figure 2: Results for the lexical decision task, Experiment 1. The vertical axis plots mean endorsement rate as a percentage with standard error bars, and the horizontal axis divides the novel words according to their phonotactic violation profile, together with an illustrative example of that profile type.

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

5.4. Results II: Is the cumulativity sub-linear, linear, or super-linear?

Since the previous analysis finds that participants inferred ganging cumulativity between constraint violations, as predicted by HG, we can further probe the nature of this effect. Is the effect of multiple simultaneous violations linearly additive (doubly-violating words are judged as marked *in proportion* to the sum of their single violations), sub-linear (doubly-violating words are *less marked* than the penalty associated with their isolated violations), or super-linear (doubly-violating words are *more marked* than the sum

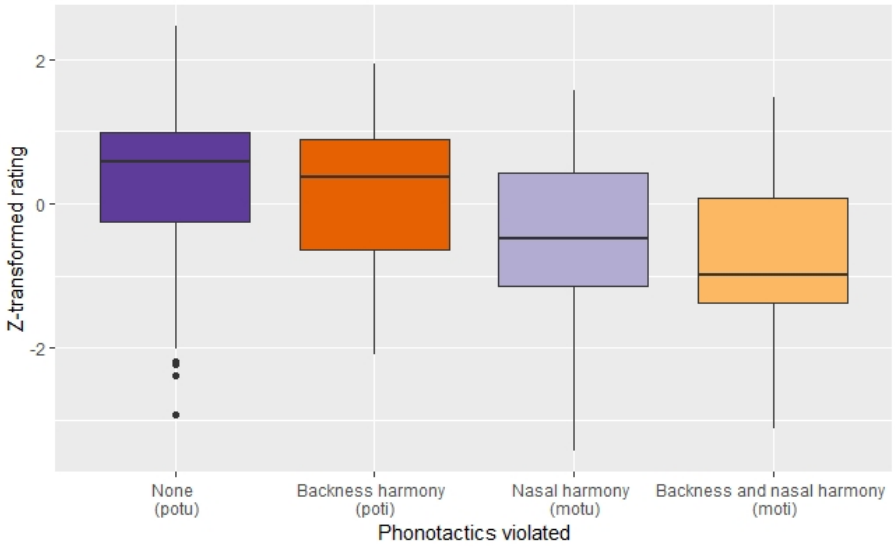

Figure 3: Results for the ratings task, Experiment 1. For readability, I plot z-normalized rating on the vertical axis, and the horizontal axis divides the novel words according to their phonotactic violation profile, together with an illustrative example of that profile type.

of their violations)? Unlike the question of *whether* constraint cumulativity exists in the general case, there is no clear consensus around the linearity of such effects in the theoretical literature. Although the original formulation of HG (Legendre et al., 1990) predicts only linearly-additive cumulativity, other frameworks that use weighted constraints make different predictions about the specific conditions under which linearity (or lack thereof) should be observed (cf. Smith and Pater (2017)). Since this topic is the subject of ongoing investigation, the linearity of any observed cumulativity will be statistically assessed so it may contribute to this literature but will not form the basis of any further theoretical claims.

To probe the linearity of the cumulativity, a second statistical analysis was carried out on the same data sets using a mixed effects regression model containing two fixed effects — whether a word violated the nasal-harmony phonotactic (*true*, *false*), whether a word violated the backness-harmony phonotactic (*true*, *false*) — and their interaction. Of interest here is the interaction term: a significant interaction with a positive coefficient would indicate that the cumulativity inferred was *sub-linear* — two coincident vi-

olations were judged less severe than would follow from the combination of their markedness when occurring in isolation. Alternatively, a significant interaction with a negative coefficient would indicate that the cumulativity was *super-linear* — two coincident violations were judged more severe than would follow from their independent status in the grammar. Finally, a non-significant interaction would be consistent with *linear* cumulativity: doubly-violating forms are judged ill-formed in proportion to the penalty associated with their component violations.

5.4.1. Lexical decision task

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

5.4.2. Ratings task

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

5.5. Local discussion

Experiment 1 provides evidence that in a poverty-of-the-stimulus environment learners infer ganging cumulativity between violations of two separate phonotactic constraints in their learning data. In the lexical decision task, doubly-violating forms were endorsed in proportion to the likelihood of endorsement of forms bearing each of their violations independently, and in the ratings task words received a rating proportional to the summed penalty of their independent violations, indicating that the cumulativity inferred is linear.

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. 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 I: Do subjects infer ganging cumulativity?

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 and a four-level fixed effect for violation profile. The log-odds of endorsement for forms violating only sibilant harmony were significantly lower than that of fully-conforming forms ($\beta = -0.968$, $p < 0.001$), as were the log-odds of endorsement for backness harmony-violating forms ($\beta = -1.223$, $p < 0.001$), paralleling the results from the lexical decision task in Experiment 1. Further, participants endorsed words violating both sibilant and backness harmony at lower rates than those which violated only sibilant harmony ($\beta = -0.940$, $p < 0.001$) or only backness harmony ($\beta = -0.685$, $p = 0.002$).

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, forms violating only sibilant harmony were rated significantly lower than fully-conforming forms ($\beta = -14.722$, $p < 0.001$), as were forms violating only backness harmony ($\beta = -14.429$, $p < 0.001$). Doubly-violating forms received lower ratings than those violating only sibilant harmony ($\beta = -15.438$, $p <$

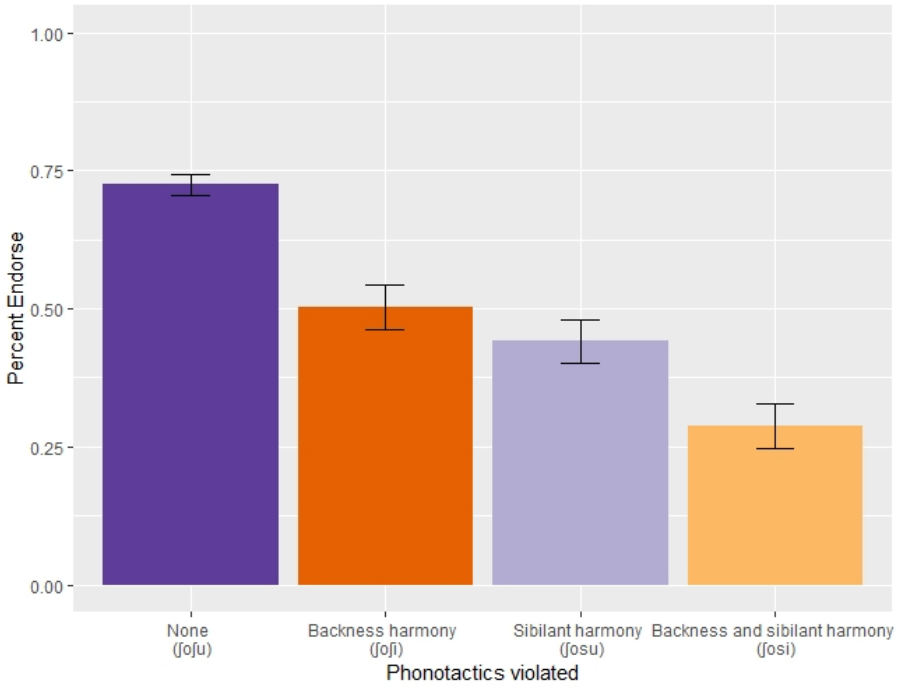

Figure 4: Results for the lexical decision task, Experiment 2.

0.001) and only backness harmony ($\beta = -15.730$, $p < 0.001$), confirming that learners inferred ganging cumulativity between distinct constraint violations.

6.3. Results II: Is the cumulativity sub-linear, linear, or super-linear?

The ganging cumulativity speakers exhibited was linear for both lexical decision (sibilant-harmony violation = *true*: $\beta = -1.003$, $p < 0.001$; backness-harmony violation = *true*: $\beta = -1.285$, $p < 0.001$; interaction $\beta = 0.241$, $p = 0.427$) and ratings tasks (sibilant harmony-violation = *true*: $\beta = -14.722$, $p < 0.001$; backness harmony-violation = *true*: $\beta = -14.429$, $p < 0.001$; interaction $\beta = -1.008$, $p = 0.790$).

6.4. 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.

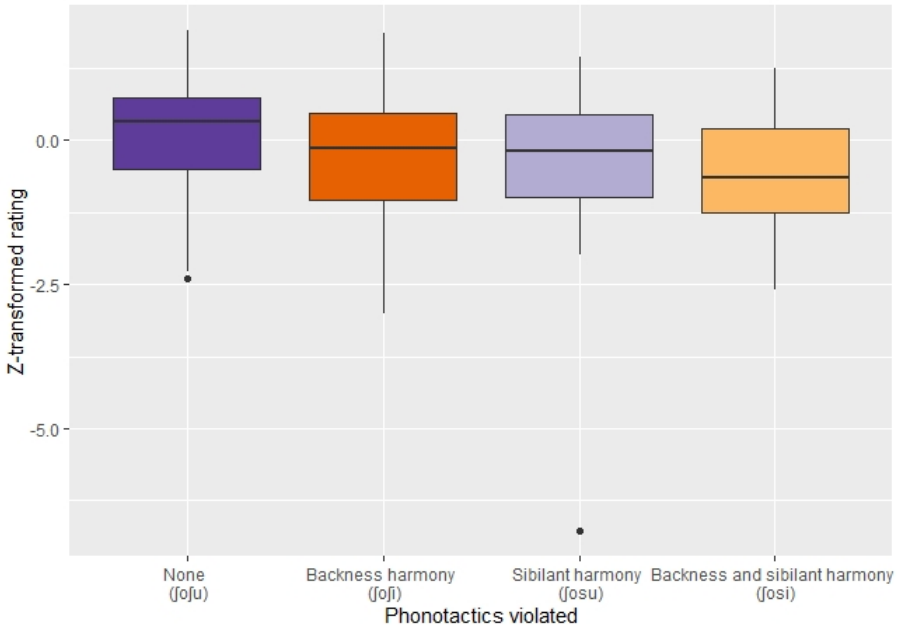

Figure 5: Results for the ratings task, Experiment 2.

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 phase items which contrasted in vowel harmony only than those which contrasted in consonant harmony only.[1] 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 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 shows the results of the lexical decision task in Experiment 3a. The final logistic regression model contained a random intercept for subject and word, and a four-level fixed effect of violation profile. Forms violating only backness harmony had significantly lower log-odds of endorsement compared to fully-conforming forms ($\beta = -0.504$, $p = 0.026$), as did those which violated only nasal harmony ($\beta = -1.562$, $p < 0.001$). Forms violating both backness harmony and nasal harmony had significantly lower log-odds

[1] 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, with the result that participants' number of errors for each type of verification trial could differ by at most 3.

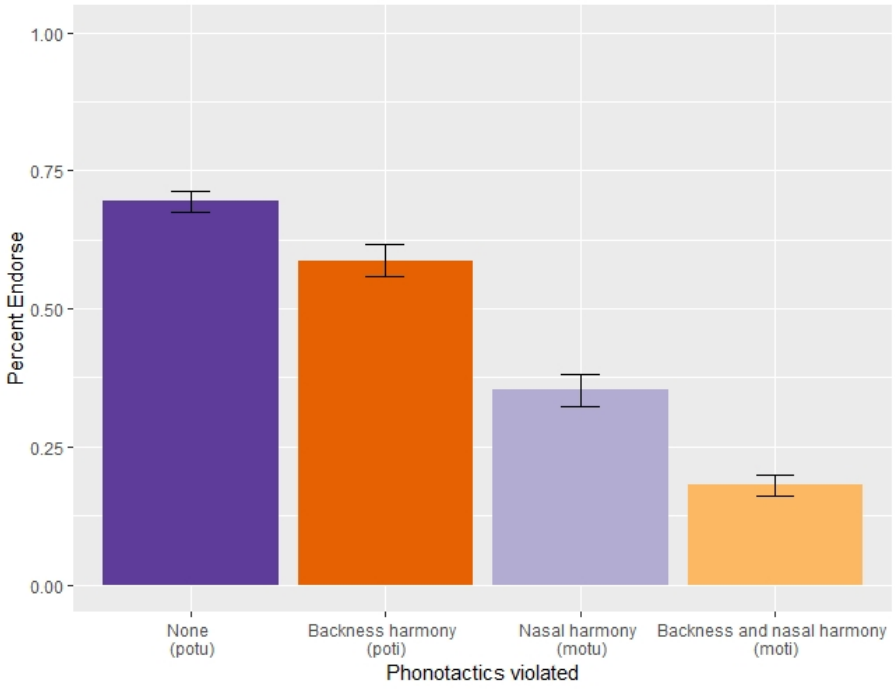

Figure 6: Results for the lexical decision task, Experiment 3a.

of endorsement than those violating only backness harmony ($\beta = -2.031$, $p < 0.001$) and those violating only nasal harmony ($\beta = -0.973$, $p < 0.001$).

7.2.2. Ratings task

Results of the ratings task are presented in Figure 7. Here, only nasal harmony-violating forms were rated significantly lower than fully-conforming word ($\beta = -17.536$, $p < 0.001$); backness harmony-violating forms did not differ significantly in their ratings from fully-conforming words ($\beta = 0.706$, $p = 0.850$). Further, doubly-violating words did not differ significantly from nasal harmony-violating words ($\beta = -7.960$, $p = 0.086$), although they did differ from backness harmony-violating words ($\beta = -26.202$, $p < 0.001$). Since the first analysis did not reveal that learners reliably rated backness-violating forms worse than fully-conforming forms — and thus did *not* exhibit constraint cumulativity — the test for linearity of ganging cumulativity was not carried out on the ratings data.

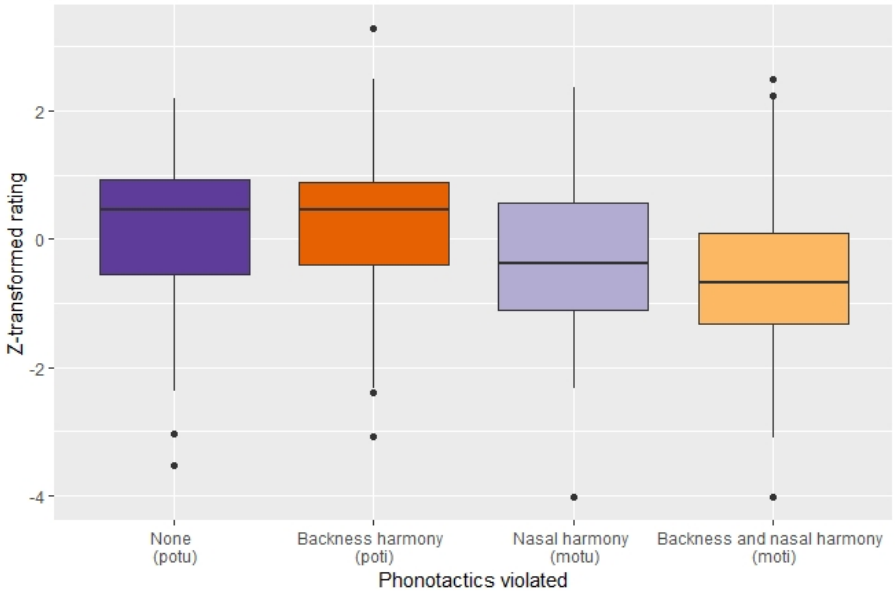

Figure 7: Results for the ratings task, Experiment 3a.

7.3. Results II: Is the cumulativity sub-linear, linear, or super-linear?

7.3.1. Lexical decision task

As before, I analyzed the log-odds of endorsing a test item in a logistic regression model with a random intercept for subject and word, and three fixed effects: nasal harmony violation (*true*, *false*), backness harmony-violation (*true*, *false*) and their interaction. I found a significant main effect of nasal harmony violation ($\beta = -1.562$, $p < 0.001$), and a significant main effect of backness harmony violation ($\beta = -0.504$, $p = 026$); the interaction term was not significant ($\beta = -0.469$, $p = 0.196$). This indicates that the cumulativity inferred did not depart from what would be expected under the linear addition of penalties for each violation.

7.4. 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 and linearly to likelihood of endorsement. In the ratings task, however, only violations of the nasal harmony phonotactic contributed to lower ratings on average. What could be the

cause of this apparent case of learners failing to infer cumulativity of constraint violations? I hypothesize that this effect is due to the change in training paradigm, and concomitant adjustment of verification standards. Passive exposure training may have resulted in more variable learning outcomes, and reduced strength of learning overall, which could have contributed to the quicker decay for the backness harmony phonotactic compared to the nasal harmony phonotactic over the course of the two generalization tasks. Anecdotal support for this hypothesis comes from the higher accuracy in the verification phase on nasal harmony violating words (mean accuracy 86.7%) compared to backness harmony violating words (mean accuracy 76.1%).

Experiment 3b

To test the post-hoc hypothesis that the lack of cumulativity in the ratings task of Experiment 3a was a task effect, a shortened, ratings-only version of the same experiment was carried out. If participants truly fail to infer constraint cumulativity under more naturalistic training conditions, the results on the ratings task in Experiment 3b should mirror those of Experiment 3a. However, if the results of the ratings task in Experiment 3a were simply due to interference from intervening time and exposure to novel nonconforming words, the results in Experiment 3b should mirror the results of ratings tasks from Experiments 1 and 2.

7.5. Methods

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

7.6. 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, forms which violated only backness harmony received lower ratings than fully-conforming forms ($\beta = -10.136$, $p = 0.012$), as did forms which violated only nasal harmony ($\beta =$

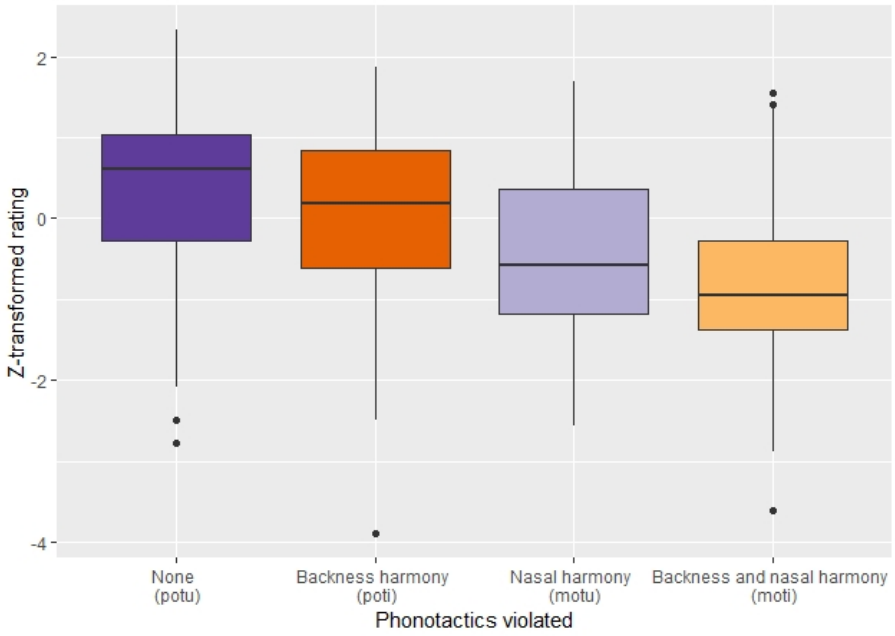

Figure 8: Results for the ratings task, Experiment 3b.

-27.029, $p < 0.001$). Further, doubly-violating forms received lower ratings than those which violated backness harmony only ($\beta = -26.725$, $p < 0.001$), as well as those which violated nasal harmony only ($\beta = -9.832$, $p = 0.044$), confirming that learners inferred ganging cumulativity between constraint violations.

7.7. Results II: Is the cumulativity sub-linear, linear, or super-linear?

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

7.8. Local discussion

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

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 allowed us to test whether counting and ganging cumulativity obtained simultaneously, since we 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 the same as for Experiment 1. 48 novel test words were created for this experiment, each four syllables long with ˈCVCVˌCVCV syllable structure and a left-aligned trochaic stress pattern (ex., *minemeni*, *putotupo*, *petipite*, etc.). 24 of these words conformed to both nasal harmony and backness harmony phonotactics, and the remaining

24 were divided evenly among the two violation levels (one locus of violation vs. two) of both phonotactics.[2]

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, 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.

8.2. Analysis

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

8.3. Results I: Did learners infer counting cumulativity?

8.3.1. Lexical decision task

Figure 9 shows the results of the lexical decision task in Experiment 4. The final logistic regression model contained a random intercept for subject and word and a three-level fixed effect of violation profile. The model indicated that, holding the level of violations of nasal harmony constant at its average value, a nonword which violated the backness harmony phonotactic in one location was less likely to be endorsed than one which did not ($\beta = -1.77$, $p < 0.001$), and a form that violated the backness harmony phonotactic in two locations was even less likely to be endorsed than one which violated it only once ($\beta = -0.548$, $p < 0.001$). For nasal harmony the complementary was true: a single violation of nasal harmony decreased the log-odds of endorsement significantly ($\beta = -0.623$, $p < 0.001$), and an increased level of

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

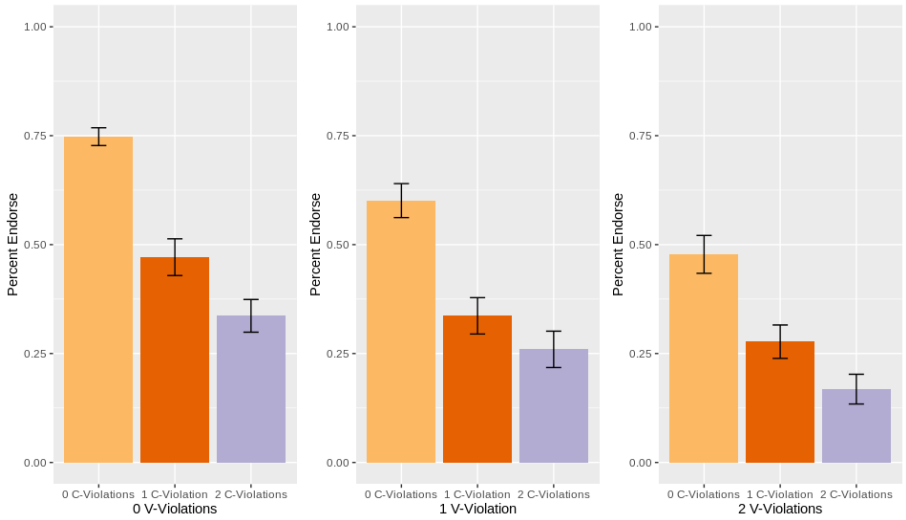

Figure 9: Results for the lexical decision task, Experiment 4. The vertical axis plots mean endorsement rate as a percentage, with standard error bars, and the horizontal axis divides the novel words according to their level of vowel-harmony violations, grouping by level of consonant violations. Note: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

violation of the same type also significantly reduced the log-odds of endorsement ($\beta = -0.473$, $p = 0.002$), with the level of violations of vowel-harmony held at its average value.

8.3.2. Ratings task

Figure 10 shows the results of the ratings task in Experiment 4. The final linear regression model contained a random intercept for subject and word and two three-level fixed effect of violation level (zero, one, or two loci of violation). For nasal harmony, holding the violation level of backness harmony constant, a single locus of violation of nasal harmony resulted in a significantly lower rating ($\beta = -6.379$, $p < 0.001$), and the further addition of a second nasal harmony violation locus resulted in a significant decrease in rating compared to words with only one nasal harmony violation ($\beta = -4.693$, $p = 0.002$). The model also indicated that, holding nasal-violation level constant, a novel word which violated the backness harmony phonotactic once did not receive a significantly lower rating than one which did not ($\beta = -0.505$, $p = 0.696$), but that a form that violated the backness harmony phonotactic twice received significantly lower ratings than a word which violated it only

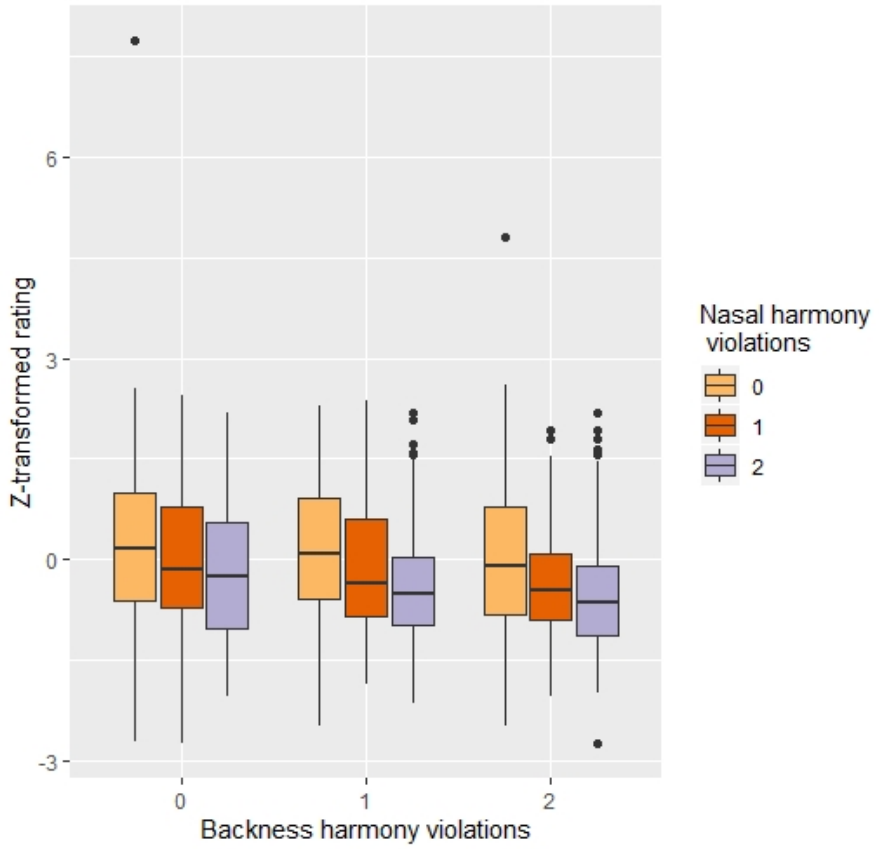

Figure 10: Results for the ratings task, Experiment 4.

once ($\beta = -4.439$, $p = 0.004$).

8.4. Local discussion

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

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 demonstrate statistically equivalent knowledge of both phonotactics in the verification phase. Another possibility is that the penalty for violating a given phonotactic 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 (formally, Δ [t . . . n], [n . . . n] $>$ Δ [o . . . e], [e . . . e], Δ [s . . . ʃ], [ʃ. . . ʃ], where Δ is a function that returns the perceptual distance between two phones). The lower perceptual distinctness of sibilant harmony-violations and backness harmony-violations relative to nasal-harmony violations could also be the cause of 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. 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 theories of phonological acquisition which hold that both input statistics and perceptual similarity play a role in shaping the grammar (Steriade, 2001; Wilson, 2006; Zuraw, 2013, *i. a.*).

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 a range of phonological models.[3] Each of these models was evaluated in the following way: first the model was exposed to the 32 two-syllable fully-conforming training data which participants were trained on in Experiments 1 and 4, and the two free parameters of the model (discussed in section **??**) were set to optimize fit to the participants' generalization data from Experiment 1. Then each model was used to predict a probability of endorsement for the four-syllable stimuli from the generalization phase of Experiment 4, scaled using a free *temperature* parameter (T) to allow for comparison between predicted probabilities and experimental results. Models were evaluated on their ability to mimic the participants' generalization data from Experiment 4, based on the learning data given to participants and the best-fitting parameter values from Experiment 1. To preview the results, all three 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, and with fewer linking hypotheses, than the other models.

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 to yield a probability distribution over four outcomes which sum to 1.

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.

[3] 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 (*in preparation*)), I address lexical decision data so as to minimize unnecessary assumptions.

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.

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], [mime], [memi], [meme], etc.

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.

10.2. Free parameters for inaccurate learning and phonetic bias

??

The human results in Table 2 depart from optimal performance in two ways: the first is that subjects learned the generalizations in their training data *inaccurately*, such that they accept phonotactically violating forms at

nonzero rates. Secondly, they exhibit an asymmetry in severity between violations of backness harmony vs. violations of nasal harmony. To allow for a closer fit to the experimental data, I equip each model with two "free parameters", one governing the *accuracy* with which the learned grammar reflects the (categorical) distribution of training stimuli in its generalization, and the other governing the role of*phonetic substance* which I have argued drives the persistent difference in severity between violations of nasal- and backness-harmony. These free parameters are incorporated into each model differently, and are optimized for each model to best fit the generalization data from Experiment 1. This process can be thought of as "solving for" the most probable learner-internal values of these parameters, conditioned on the observed training data, the structure of the theory being evaluated, and the attested generalization data. These optimal values are then used to make predictions about generalization data from Experiment 4, which will be used to evaluate the quality of each model.

10.3. Traditional approaches to phonotactic learning, and the Multiple Grammars model

I first evaluated the Recursive Constraint Demotion algorithm (RCD; Tesar and Smolensky (2000)).[4] 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 first installing only winner-preferring constraints in the highest stratum possible. Then, if there are no unranked constraints that prefer only winners, RCD chooses a constraint to install in the next stratum at random, and proceeds until there are no unranked constraints.

The algorithm was provided with *mark-data pairs* consistent with the 32 fully-conforming stimuli from Experiment 1. Constraints ranked were AGREE([BACK]), AGREE([NASAL]), ID-[BACK], and ID-[NASAL]; learning simulations were carried out in OTSoft (Hayes et al., 2003). The algorithm returned the following pairwise rankings: AGREE([BACK]) >> Id([back]), AGREE([NASAL]) >> ID-[NASAL], which implies the stratification below:

[4]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.

Stratum 1: AGREE([BACK]), AGREE([NASAL])
Stratum 2: ID-[BACK], ID-[NASAL]

As designed, RCD learned a Markedness >> Faithfulness ranking which is maximally restrictive (Tesar and Smolensky, 2000), and admits only fully-conforming items when generalizing to new forms. Though the model performed as intended, the algorithm was not designed with gradient phonotactics in mind, and thus exhibited is no trace of the cumulative behavior that human learners did: a singly-violating form such as *mipi* is not distinguished from doubly-violating *mipu*. Further, violations of nasal harmony and violations of backness harmony are treated identically by the model. Thus, on its own, the RCD algorithm is incapable of approaching testability against the experimental data: a linking hypothesis is needed. I adopt one in which the possibility of inaccurate perception of the learning data allows the learner to consider a distribution over possible discrete RCD-based grammars: on each trial there is a constant small probability of misperceiving a consonant or vowel. From this noisy input the learner learns not a single probabilistic grammar but a distribution over categorical grammars which are consistent with the learning data, weighted by the likelihood of having perceived that data supporting that grammar. I term this augmented model the Multiple Grammars model.

10.3.1. Setting up the Multiple Grammars model

The core RCD-based algorithm is initialized with the constraints AGREE([BACK]), AGREE([NASAL]), ID-[BACK], and ID-[NASAL]. Before exposure to training data, the learner has a categorical Markedness >> Faithfulness ranking of the four relevant constraints. On each of the 32 learning trials, the learner updates that ranking probability based on the distribution over phonological forms that trial provides evidence for. During generalization, a similar process occurs: when presented with a form to be judged, the learner "perceives" that form as a probability distribution over possible phonological forms. The probability of acceptance for each novel stimulus is equal to the summed probability of accepting each of the possible realizations of that stimulus, multiplied by the probability of the grammar's ranking being such that that stimulus is accepted. Aggregate results for each stimulus type are obtained by averaging over all items in that class.

This procedure is illustrated with the following example. Given that the probability of mishearing a consonant = probability of mishearing a vowel

= 10%, a waveform *mimi* is perceived as the distribution over SRs shown in Table 4.

Waveform	Stimulus profile	Candidate SR	Prob. C_1 accurate	Prob. V_1 accurate	Prob. C_2 accurate	Prob. V_2 accurate	Prob. of UR given waveform	Total prob. of profile
mimi	*Fully-conforming*	[mimi]	0.9 ×	0.9 ×	0.9 ×	0.9 =	0.6561	0.6724
	Fully-conforming	[pipi]	0.1	0.9	0.1	0.9	0.0081	
	Fully-conforming	[mumu]	0.9	0.1	0.9	0.1	0.0081	
	Fully-conforming	[pupu]	0.1	0.1	0.1	0.1	0.0001	
	Backness-violating	[mimu]	0.9	0.9	0.9	0.1	0.0729	0.1476
	Backness-violating	[pipu]	0.1	0.9	0.1	0.1	0.0009	
	Backness-violating	[mumi]	0.9	0.1	0.9	0.9	0.0729	
	Backness-violating	[pupi]	0.1	0.1	0.1	0.9	0.0009	
	Nasal-violating	[mipi]	0.9	0.9	0.1	0.9	0.0729	0.1476
	Nasal-violating	[pimi]	0.1	0.9	0.9	0.9	0.0729	
	Nasal-violating	[mupu]	0.9	0.1	0.1	0.1	0.0009	
	Nasal-violating	[pumu]	0.1	0.1	0.9	0.1	0.0009	
	Doubly-violating	[mipu]	0.9	0.9	0.1	0.1	0.0081	0.0324
	Doubly-violating	[pimu]	0.1	0.9	0.9	0.1	0.0081	
	Doubly-violating	[mupi]	0.9	0.1	0.1	0.9	0.0081	
	Doubly-violating	[pumi]	0.1	0.1	0.9	0.9	0.0081	

Table 4: A schematic example of the linking hypothesis in the Multiple Grammars model, with the waveform *mimi* being perceived as a distribution over SRs.

The distribution over possible constraint rankings — and thus possible grammars — is updated in accordance to the perceived distribution over words. Since the total probability mass on possible words which violate AGREE([BACK]) is 0.0324 + 0.1470 = 0.18, the probability that this constraint is ranked above ID-[BACK] after exposure to *mimi* is (1-0.18) = 0.82;

the same is true of AGREE([NASAL]) and ID-[NASAL].

If, after exposure to just this one learning datum, the model were asked to judge the novel waveform *pupu*, it would proceed in the following way.[5] Via the same calculation as above, the probability that *pupu* is perceived as violating AGREE([BACK]) is 0.18, and of violating AGREE([NASAL]) is 0.18. Since the probability of a grammar accepting a nasal-violating form is 0.18 and of accepting a backness-violating form is 0.18, the model predicts that the novel form will be accepted 70.48% of the time ((1−(p(AGREE([NASAL]) >> ID-[NASAL])) × p(violate AGREE([NASAL]))) + (1−(AGREE([BACK]) >> ID-[BACK]) × p(violate AGREE([BACK])))), and rejected (1− 0.7048) = 29.52% percent of the time.

10.3.2. Fitting the Multiple Grammars model

The values of the misperception parameters which best fit the results of Experiment 1 were as follows: $p_{misperceiveC} = 0.8\%$, $p_{misperceiveV} = 1.4\%$. Predictions for the experimental data from Experiment 1 is in Table 5.

Violations	Exp. 1 (scaled)	Multiple Grammars (scaled)
None	0.39	0.45
Backness	0.30	0.27
Nasal	0.19	0.18
Backness & nasal	0.11	0.10

Table 5: Multiple Grammars numerical model fit to Experiment 1 data, scaled and unscaled.

It is clear that the model can qualitatively distinguish four different levels of acceptance, which correspond to the four aggregate rates for the generalization data in Experiment 1.

10.4. 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 comes equipped with a range of a range of phonological learning algorithms (Boersma and Hayes, 2001; Jäger, 2007; Magri, 2014;

[5]Note that when modeling Experiment 1 this process was repeated a further 31 times, omitted here for the sake of clarity.

Magri and Storme, 2018, among others). It employs an error-driven update rule to learn ranking values for each constraint, which serve as the mean of a Gaussian distribution from which individual rankings are sampled when the grammar evaluates novel forms.

10.4.1. 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 — 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 (for precedent see Hale and Reiss (1998)).

In principle, the full input to such a model would be all possible CVCV forms with C ∈ {/p, m/} and V ∈ {/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 have not begun to approach the Faithfulness constraints on the ranking-value continuum, the algorithm mistakenly allows disharmonic URs such as /pimi/ to realize without repair, as [pimi]. However, if the algorithm is allowed to run its course, it will promote Markedness

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 begin to 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/ $\rightarrow$ [pimi $\sim$ pipi]. It is in this zone where non-categorical outcomes are permitted that the possibility for modeling cumulative constraint interaction lies.

10.4.2. Fitting the Stochastic OT model

The two Faithfulness constraints were initialized with ranking value 100, and the Markedness constraints with ranking value 0. 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, I tested the model at a range of fixed ranking values which fell within the intermediary zone of overlapping Gaussians discussed above. The *accuracy* free parameter was implemented by manipulating the difference in ranking values between Markedness and Faithfulness constraints, since it is this difference which governs to what degree the constraints can interact and potentially exhibit cumulative behavior. Similarly, the *phonetic substance* free parameter was manipulated by varying 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. Both these values were fit by hand-adjusting the constraint weights down to a tenth of a point of ranking value, and testing the fit to the experimental data. 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 6.

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

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

10.5. 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.

10.5.1. Setting up the MaxEnt model

The *accuracy* free parameter was implemented by placing a Gaussian prior with mean (μ) = 0 and narrow standard deviation (σ) on the weights of AGREE([BACK]) and AGREE([NASAL]). This encourages the model to prefer solutions which have lower constraint weights in general, and thus forces 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. The *phonetic substance* parameter was instantiated with differing values of σ 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.

10.5.2. 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 σ which best fit the experimental data were $\sigma_{\text{AGREE}([\text{NASAL}])} = 0.22$, and $\sigma_{\text{AGREE}([\text{BACK}])} = 0.15$; model predictions for the Experiment 1 generalization phase data are shown in Table 7 below.

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 7: 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, we have demonstrated that all three models can qualitatively capture the ganging cumulativity observed in the generalization phase of Experiment 1. I now turn to evaluate the models on their ability to match human generalization on the data from Experiment 4, using the optimal parameters learned from Experiment 1.

11.1. The data to be modeled: Experiment 4 revisited

The unscaled generalization results from Experiment 4 are reproduced below in Table 8. 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. For example, the novel word *mitepuni* incurred 3 violations of AGREE[BACK] ([i. . . . . u, e. . . u,

Violation profile	Exp. 4
0C, 0V	0.74
0C, 3V	0.60
3C, 0V	0.47
3C, 3V	0.34
0C, 4V	0.48
3C, 4V	0.28
4C, 0V	0.34
4C, 3V	0.26
4C, 4V	0.17

Table 8: Lexical decision data from Experiment 4. Note that C and V in the leftmost column denote the number of violations of the consonant-harmony and vowel-harmony phonotactics for that candidate class respectively.

u. . . i]), and 4 violations of AGREE[NASAL] ([m. . . t, m. . . . . p, p. . . n, t. . . . . n]).[6] This allows the modeling to remain in line with the statistical analysis in section 8.

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. As noted earlier, to allow for a direct comparison between the mean percent endorsement in the experimental data and the predicted percent endorsement given by the models, model predictions are adjusted using a *temperature* term (T), following Smolensky (1986); Hayes and Wilson (2008). T is a free parameter set on a by-model basis to maximize the fit of the model's predictions to the data, and is used to generate adjusted model predictions in the following way: for a given form with predicted percent endorse x given by a model, the adjusted probability of endorsement for that form is $x^{1/T}$. It is this adjusted predicted percent endorse (henceforth simply *percent endorse*) that is the metric of in-

[6] 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 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.

terest for the remainder of the paper.[7] Table 9 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.

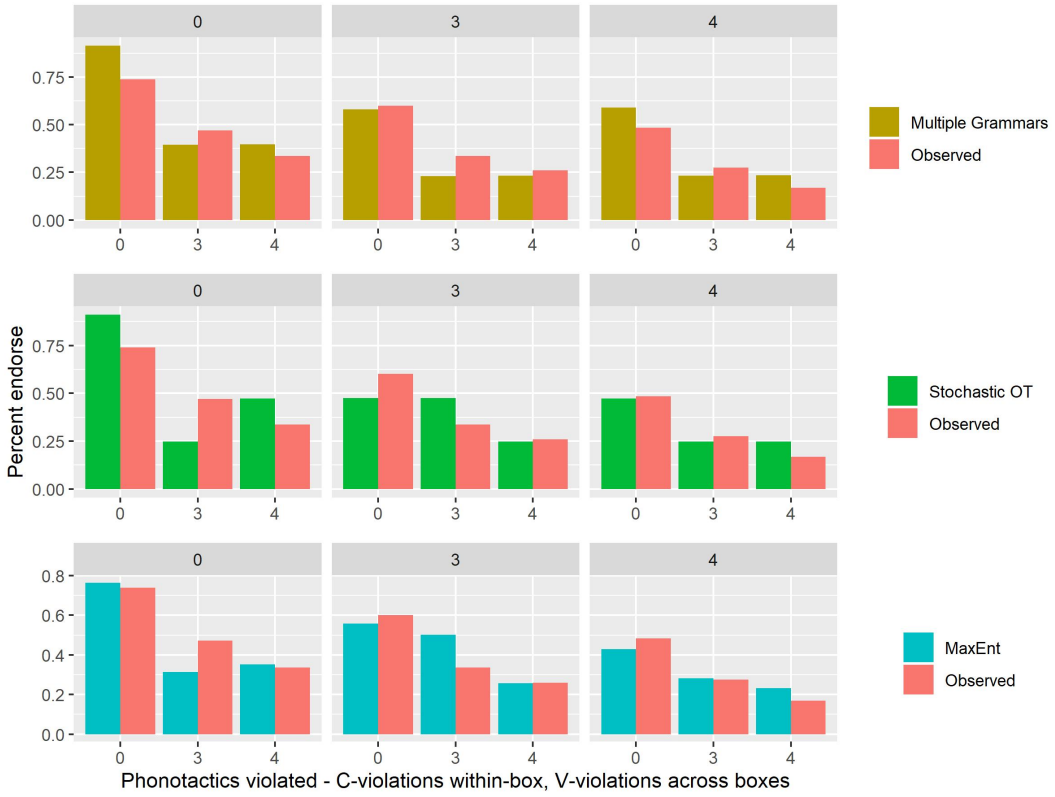

Figure 11: Multiple Grammars, Stochastic OT, and MaxEnt model predictions for Experiment 4 lexical decision data.

Model fits are evaluated by calculating the *sum of squared errors* for each of the violation profiles in Experiment 4. Log-likelihood values were not calculated because we are interested in 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 10.

[7] T was not used when fitting the models to data from Experiment 1 to minimize degrees of freedom.

Violation profile	Exp. 4	Multiple Grammars	Stochastic OT	MaxEnt
0C, 0V	0.74	0.91	0.91	0.76
0C, 3V	0.60	0.58	0.47	0.55
3C, 0V	0.47	0.40	0.47	0.31
3C, 3V	0.34	0.23	0.25	0.50
0C, 4V	0.48	0.59	0.47	0.42
3C, 4V	0.28	0.23	0.25	0.28
4C, 0V	0.34	0.40	0.47	0.35
4C, 3V	0.26	0.23	0.25	0.26
4C, 4V	0.17	0.23	0.25	0.23

Table 9: Multiple Grammars, 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.

Model	Sum of squared error
Multiple Grammars	0.069
Stochastic OT	0.138
MaxEnt	0.062

Table 10: Model comparison metrics using distance-agnostic AGREE.

At a glance, we can see that the models vary in their ability to match human generalization to the novel data. While all models can capture ganging 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, as does the Multiple Grammars model, though the effect is obscured by the uncertainty introduced by the possibility of inaccurately-perceived test items. 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 Multiple Grammars and 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 best; Stochastic OT predicts a poorly-fitting distribution over output forms, while the Multiple Grammars model comes considerably closer to the mark.

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 11.

Model	Sum of squared error
Multiple Grammars	0.089
Stochastic OT	0.093
MaxEnt	0.085

Table 11: 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. This effect was linearly additive: doubly-violating words were judged ill-formed in proportion to the summed ill-formedness associated with each of their phonotactic violations. Experiment 2 replicated these findings using a different combination of phonotactics, sibilant-harmony and backness harmony, and Experiments 3a and 3b again replicated 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, alongside ganging cumulativity. 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 default, 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 several 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 three models were able to meet. Turning to these models, I compared the ability of a Multiple Grammars model (a RCD-based algorithm equipped with a rich perceptually-based linking hypothesis), a Stochastic OT model, and a MaxEnt model to match the generalization behavior of participants in Experiment 4. Fitting the models' free parameters to data from 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. The results indicated that while all models were able to capture the ganging cumulativity in Experiments 1 and 4, the MaxEnt model was best able to capture both the counting and ganging cumulativity participants exhibited in Experiment 4, and did so without the additional linking hypothesis which was needed to support the Multiple Grammars model.

13. Conclusion

Taken together, the experimental data and the modeling thereof indicate that both counting and ganging constraint cumulativity are default in 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 receive the most support from the present results.

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.

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

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.

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.

Hale, M. and Reiss, C. (1998). Formal and empirical arguments concerning phonological acquisition. *Linguistic inquiry*, 29(4):656–683.

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. (2003). 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.

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.

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

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.

Lionnet, F. A. (2016). Subphonemic teamwork: A typology and theory of cumulative coarticulatory effects in phonology.

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*.

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. (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.

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.

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.

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.