Phonology **37** (2020) 551–576. © The Author(s), 2021. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/S0952675720000275

*Constraint cumulativity in phonotactics: evidence from artificial grammar learning studies**

Canaan Breiss
University of California, Los Angeles

An ongoing debate in phonology concerns the treatment of cumulative constraint interactions, or 'gang effects', and by extension 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 cumulative constraint violations in phonotactics in the absence of a confounding natural-language lexicon. I find that learners consistently infer linear counting and ganging cumulativity across a range of phonotactic violations.

1 Introduction

The treatment of CUMULATIVE CONSTRAINT INTERACTIONS is the subject of ongoing debate in phonology. In this paper I take on the topic of cumulative constraint interactions in phonotactics (Albright 2009, 2012, Pizzo 2015, Durvasula & Liter 2020). I use a method inspired by work in experimental syntax (Featherston 2005, 2019) in which syntactic violations are manipulated in a crossed experimental design in order to tease apart the independent contribution of each one, and to gain insight into how multiple violations are combined in the grammar. I combine this independent

* E-mail: CBREISS@UCLA.EDU.
Many thanks to Bruce Hayes and Megha Sundara for their guidance and advice on all aspects of this research. Thanks also to Adam Albright, Adam Chong, Karthik Durvasula, Sara Finley, Shigeto Kawahara, James White, Colin Wilson, two anonymous reviewers for *Phonology* and members of the audiences at the LSA Annual Meeting 2019, SCAMP 2019 and the UCLA Phonology seminar for much helpful commentary and discussion. I am also grateful to Beth Sturman and Eleanor Glewwe for recording stimuli, to Henry Tehrani for technical assistance and to Ruby Dennis, Shreya Donepudi, Grace Hon, Lakenya Riley, Azadeh Safakish and Amanda Singleton 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.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

manipulation of violations with an artificial grammar learning paradigm which imposes a 'sandbox' environment on the learner, where the statistics of the language being learned can be carefully controlled. Doing this ensures that whatever generalisations participants form about the (non-) interaction of independent phonotactic violations can be taken to reflect properties of the structure of the grammar that is learned, rather than asymmetrical distributions of structures in the lexicon; this point is taken up again in §3.3. Note that the use of an artificial language does not render the experimental results impervious to the influence of whatever non-linguistic cognitive factors may be at play in acceptability judgements. We expect such effects (though I do not model them here explicitly), but do not anticipate that they will exert an asymmetrical effect on different items in the experiments, so the within-experiment comparisons which are the focus of this paper should be unbiased.

I show that learners consistently infer linear COUNTING and GANGING cumulativity across a range of phonotactic violations. I discuss the compatibility of these results with a range of contemporary constraint-based phonological frameworks, and argue that only probabilistic weighted-constraint frameworks such as Maximum Entropy Harmonic Grammar (Smolensky 1986, Goldwater & Johnson 2003) and Noisy Harmonic Grammar (Boersma & Pater 2016) are able to capture both counting and ganging cumulativity.

2 Constraint cumulativity in phonological theory

Constraint-based phonological frameworks diverge on whether they can model cumulative constraint interactions. Classical Optimality Theory (OT; Prince & Smolensky 1993) 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) holds that speakers take an informationally holistic approach, considering all constraint violations when choosing the optimal outcome. The difference can be observed in the schematic tableaux in (1).

(1) a.

	CONSTRAINT A	CONSTRAINT B
Candidate 1	*!	
☞ Candidate 2		**

b.

	CONSTRAINT A $w = 3$	CONSTRAINT B $w = 2$	$\mathcal{H}$
☞ Candidate 1	*		3
Candidate 2		**	4

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

In OT in (1a), Candidate 2 wins out at the expense of Candidate 1, because Candidate 1 violates the higher-ranked Constraint A, while Candidate 2 does not. Because Constraint A is ranked above Constraint B, Candidate 1's single violation of Constraint A is more important than Candidate 2's two violations of Constraint B. This removes Candidate 1 from contention, leaving Candidate 2 as optimal. In HG in (1b), the optimal outcome is the one which has the lowest harmony penalty when *all* violations are considered. Each candidate's harmony is equal to the number of times it violates each constraint, multiplied by the weight (w) of the constraint violated. Using this method, the same violations result in Candidate 1 being optimal, because it has a lower harmony than Candidate 2.[1] This is because the two violations of Constraint B, though tolerated individually, together outweigh the penalty associated with the single violation of Constraint A. Thus HG and OT sometimes predict different outcomes from the same schematic example, because candidates' violations are cumulative in HG, but not in OT. Although constraint violations in OT can be compared numerically when two candidates tie on all higher-ranked constraints, these are not considered cases of cumulative constraint interaction. On the other hand, HG *cannot help* but exhibit cumulativity, regardless of the relative strengths of the constraints involved.

Jäger & Rosenbach (2006) identify two possible types of constraint cumulativity: COUNTING CUMULATIVITY and GANGING CUMULATIVITY. Counting cumulativity, illustrated above, occurs when one violation of a lower-weighted constraint leads to a lower penalty than one violation of a higher-weighted constraint, but two or more violations of the first constraint together assign a higher penalty than a single violation of the second. Ganging cumulativity is found when independent violations of low-weighted constraints assign a lower penalty than a single violation of a higher-weighted constraint, but, when they occur together, these lower-weighted violations 'gang up' together to yield a more severe penalty.

3 Constraint cumulativity in phonological typology

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

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

[2] Note, however, that Local Constraint Conjunction and HG are not necessarily incompatible (cf. Shih 2017), but the role that Local Constraint Conjunction

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

3.1 Constraint cumulativity in alternations

Evidence for constraint cumulativity is often discussed in the context of conditions on the relationship between phonological inputs and outputs, i.e. alternations (e.g. Goldwater & Johnson 2003, Coetzee & Pater 2006, Pater 2009, Zuraw & Hayes 2017). The majority of phonological patterns which are suggestive of a cumulative analysis are cases of ganging cumulativity, where two (or more) distinct factors interact to determine how the SR for a given UR is realised. For instance, the cumulative combination of factors influencing the likelihood of *-t*/*-d*-deletion in corpus data was pointed out as a difficulty for OT by Guy (1997). Rose & King (2007) used a speech-error elicitation task to examine the effect of the simultaneous violation of various consonant co-occurrence restrictions in two Ethiopian Semitic languages, Chaha and Amharic. They found that participants produced more errors when stimuli violated several constraints simultaneously than when they violated each constraint independently. Pater (2009) analyses data from Japanese loanwords (from Nishimura 2003) to argue that the static phonotactic restriction known as Lyman's Law, which prohibits multiple voiced obstruents within a word, can be construed as a case of constraint cumulativity.[3] Pater notes that, while speakers tolerate unrepaired voiced obstruents and geminate consonants when adapting loanwords, they prefer to repair words which contain voiced geminates by devoicing them. Kawahara (2011a, b, 2013) tests this formal analysis with a series of acceptability-judgement studies, and finds robust support for Pater's conclusions. Kawahara (2012) also finds experimental evidence that Lyman's Law violations can block a voicing alternation in the native Japanese lexicon known as *rendaku*, which is triggered by compound formation (this is a further case of apparent cumulativity, supporting observations made by Itô & Mester 1986). Studies by Kawahara (2020) and Kawahara & Breiss (to appear) also indicate that the relationship between form and meaning characteristic of sound-symbolism displays cumulative effects. There has been less work on the counting cumulativity front, though recent findings by Kim (2019) demonstrate that two nasals are required to block *rendaku* application in Japanese compounds; one is not enough.

3.2 Constraint cumulativity in phonotactics

Data which suggest cumulative constraint interaction have also been noted in phonotactics, generally taking the form of additive effects of multiple marked structures on the likelihood of lexical attestation or experimentally assessed acceptability. Durvasula & Liter (2020) used an artificial grammar

plays in HG is quite different from how it is used in OT, and it will therefore not be discussed further here.

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

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

learning paradigm to examine the kinds of generalisations that were formed during phonotactic learning, and found that speakers learned multiple generalisations consistent with their data. Crucially, they also found that, when speakers were asked to extend these generalisations to novel items, the generalisations interacted in a cumulative manner: a novel stimulus violating two phonotactics was more likely to be rejected than a stimulus violating only one.

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

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

While suggestive of cumulative behaviour, however, the findings of Pizzo and Albright have an alternative explanation. This is because, in their studies, experimentally determined well-formedness is highly correlated with the lexical frequency of the very structures which are being judged. Even setting aside models which explicitly use the number of similar words in the lexicon to estimate acceptability (e.g. the Generalised Neighbourhood Model of Bailey & Hahn 2001), the prominent role of lexical statistics in influencing well-formedness judgements is well established (see Pierrehumbert (to appear) for an overview). Pioneering work by Coleman & Pierrehumbert (1997) highlighted the connections between the lexicon and phonotactic well-formedness in their predictive model of non-word judgements, inspiring much further work (e.g. Frisch *et al*. 2000, Shademan 2007, Daland *et al*. 2011, Jarosz & Rysling 2017). In addition, Albright (2012), Fukazawa *et al*. (2015) and Kawahara & Sano (2016) find evidence for a complex interaction between lexical statistics and phonological acceptability: underattestation of words in the lexicon which contain *two*

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

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.

Furthermore, there is evidence that the relationship between lexicon and phonology is diachronically bidirectional: Martin (2007, 2011) found that, assuming that speakers prefer to reuse novel coinages which are phonotactically more 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 judgements which are sensitive to lexical statistics and lexical statistics which are shaped by a synchronic preference for phonotactic well-formedness. Therefore the question of causality in natural languages – whether words are judged to be ill-formed because they are improbable in the context of the lexicon or whether skewed lexical statistics are the product of the phonological grammar – cannot be satisfactorily resolved. This prevents us from taking evidence of phonotactic cumulativity in natural languages, such as that in Albright (2012) and Pizzo (2015), as unbiased evidence for the nature of the grammar.

4 Experimental design

To tease apart phonotactic acceptability and lexical frequency, I used an artificial grammar learning paradigm to create a 'sandbox environment', in which lexical statistics can be carefully controlled, as in Durvasula & Liter (2020). Four experiments were carried out; training data for all four consisted of a set of 32 CVCV words, with individual consonant and vowel phonemes balanced for overall frequency and syllabic position, subject to the harmony constraints in each language, discussed below. Further, all local phoneme bigrams (e.g. adjacent CV and VC sequences) and all non-local phoneme bigrams on C or V tiers (e.g. the C…C sequences and V…V sequences in the CVCV words) were equally frequent. This allows me to interpret participants' inferences about (non-)cumulativity, made in the absence of disambiguating evidence and distributional asymmetries of the lexicon, to be revealing of the nature of phonotactic grammar.

Turning to the specific phonotactics involved, all four experiments involved paired varieties of consonant and vowel harmony. These phenomena have traditionally constituted core areas 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 artificial grammar learning experiments (e.g. Finley 2015, Lai 2015). Consonant harmony regulated all and only a word's consonants, and vowel harmony regulated the vowels, allowing a word to conform to or violate each phonotactic independently.[4] In Experiments 1, 3a, 3b and 4 I used consonant-nasality harmony (hereafter NASAL HARMONY): in conforming words, all consonants agree in nasality,

[4] In this paper I use the term 'phonotactic' as shorthand to refer to a static syntagmatic restriction on phoneme sequences in a language; a specific instance of the type of restrictions which the term 'phonotactics' references as a group.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

being drawn either from the nasal stops [m n] or from the voiceless oral stops [p t]; for a survey of parallels in natural languages, see Hansson (2010: 111). In Experiment 2 I used SIBILANT HARMONY: in conforming words, sibilant consonants in a word agree in anteriority, being drawn from [s z] or [ʃ ʒ] (see Hansson 2010: 55 for a typological survey). All experiments used vowel-backness (as well as rounding) harmony, referred to hereafter as BACKNESS HARMONY: in conforming words, all vowels in a word agreed in backness, being drawn from one of the sets [i e] and [u o] (for an overview, see Walker 2011).

To ensure an accurate assessment of participants' well-formedness judgements, I elicited acceptability judgements from participants using two different tasks. First, participants rendered categorical well-formedness judgements in what I term a 'binary decision task', in which they were asked to judge whether a novel word could belong to the language that they learned at the start of the experiment (possible answers were *yes* or *no*). They then completed a 'ratings task', in which they were asked to assign each of those same words a numerical rating on a 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 for either outcome – whether speakers display cumulativity or not – should be the result of converging evidence from the binary decision and ratings tasks.[5]

5 Experiment 1

In Experiment 1, I tested whether learners inferred a cumulative effect involving violations of two different phonotactics, an instance of ganging cumulativity.

5.1 Methods

5.1.1 *Participants.* 45 undergraduate students at the University of California, Los Angeles were recruited to participate in this experiment,

[5] An anonymous reviewer raises the possibility that participants could simply be judging the well-formedness of novel items on the basis of some non-phonological measure of similarity (such as the *n*-gram probabilities of the string), based on the items seen in the training phase, and thus not be inducing markedness constraints at all. While this is theoretically possible, since all phoneme bigrams in the generalisation items appeared in the exposure items, such a generalisation would come down to either tracking tier-based *n*-grams or tracking counts over trigram windows of the string. The degree to which these generalisations are 'non-phonological' is debatable, however, and a topic of ongoing investigation (cf. e.g. Wilson & Gallagher 2018). Here, I assume that whatever types of generalisations participants are forming are at least linguistically informed, and thus in the domain of the two generative theories tested in the paper, but I leave open the exact structure of these generalisations (though see Durvasula & Liter 2020 for work focusing on exactly what level of representational granularity learners form generalisations over in artificial grammar learning experiments). The raw data from all experiments reported here can be accessed in the online supplementary materials, available at https://doi.org/10.1017/S0952675720000275.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

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 (on which more below; $n = 10$), leaving 33 participants whose data was included in the final analysis.

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

For the verification phase I created two sets of items, each consisting of 16 pairs of minimally differing non-words. One member of each pair was a fully conforming word from the exposure phase, and the other was created by changing one of the consonants or vowels in the fully conforming word. Thus the pair of words differed only in a single instance of that segment. In each set, eight pairs differed in a violation of nasal harmony, and eight in a violation of backness harmony, with differences between members of each pair balanced for segmental placement and identity. For example, the familiar word [potu] was modified by altering the nasality specification of its second consonant, yielding the pair [potu] *vs*. [ponu].

In a pilot study, participants showed a strong preference for forms with identical consonants or vowels (in line with the general findings of Gallagher 2013), despite there not being any more training stimuli containing a given identical pair of phonemes than any other combination of phonemes. The verification trials were therefore structured to neutralise this confound: pairs whose fully conforming word had identical consonants (e.g. [totu]) differed only in their violation of backness harmony (e.g. [totu] *vs*. [toti]). In pairs whose conforming word contained identical vowels, the two words differed only in a violation of nasal harmony. Crucially, there were no doubly violating words in the verification phase: the purpose was simply to ensure that subjects had learned each of the two phonotactic constraints independently.

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

All words were recorded using PCQuirer by a phonetically trained native speaker of English. They were digitised at 44,100 Hz and normalised 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 generalisation tasks in the test phase:

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

the binary decision task and then the ratings task. The exposure phase consisted of two blocks of 32 pseudo-randomised self-paced trials in which words were presented auditorily, without feedback. During the exposure phase, similar-sounding items were presented together in four blocks of eight, with each subject assigned at random to one of four counterbalanced orders. For example, in one counterbalanced group, participants first heard eight words with front vowels and voiceless stops ([peti, tipi, tepe, piti, …]), followed by eight words with back vowels and nasal stops ([monu, nunu, mumo, numo, …]), eight words with back vowels and voiceless stops ([topu, pupo, topo, putu, …]) and 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-randomised trials in the exposure phase, after which they completed a second verification phase. The two sets of 16 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 a demographic questionnaire, and did not complete the test phase.

If subjects met the criteria on the verification phase, they moved on to the test phase, which consisted of the binary decision task and the ratings task. Both tasks used the same set of novel words. In the binary decision task, participants were presented with two repetitions of 48 novel words in a random order, and were asked to choose whether they thought each word could belong to the language they had learned. In the ratings task, participants were asked to rate each of the same words on a scale from 0 (*very bad*) to 100 (*very good*), based on how the word sounded as an example of the language they had learned. Demographic information was collected at the end of the experiment. The full experiment lasted approximately 15–20 minutes, depending on the number of exposure blocks the subject required.

5.1.4 *Procedure*. Participants were tested individually in a sound-attenuated room using a modified version of the Experigen platform (Becker & Levine 2020). After obtaining informed consent from the participants, the experiment began with participants being told that they would be learning a new language, after which they would be tested on their knowledge. Participants were encouraged to repeat 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

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

about how the language sounded and what their 'gut' told them was right, and to not overthink 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, and were instructed (via onscreen text) to repeat the word aloud. The verification phase had a similar structure, except that each trial played the pair of words in a random order, and participants were instructed to say both words aloud 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 before either making the binary decision or assigning it a rating.

5.2 Analysis

Data from the test phase was analysed with mixed-effects regression models in R (R Core Team 2020), using the *lme4* and *lmerTest* packages (Bates *et al*. 2015, Kuznetsova *et al*. 2017). For all statistical analyses, a maximally specified model was first fitted (following Barr *et al*. 2013); this contained a random intercept for subject and item, fixed effects of violation of backness harmony, violation of nasal harmony and their interaction, and random slopes for all fixed effects by subject. Dummy coding was used for the two fixed effects. In cases of non-convergence, interactions among random slopes were removed first, then the slopes themselves, until the model converged.

For the binary decision task I modelled the log-odds of endorsing an item as a function of its phonotactic violations using mixed-effects logistic regression. Note that, since this task involves a binary outcome, I examined the models for cumulativity on the log-odds scale, rather than the probability scale. For the ratings task I modelled the raw numerical data using mixed-effects linear regression.[6]

Regardless of the domain of analysis (log-odds of endorsement or numerical rating), once a model was fitted, I explored the interaction term, using planned comparisons to test for a difference between the singly violating levels and the doubly violating level using the *glht*() function from the *multcomp* package (Hothorn *et al*. 2016). Probing this difference is important, since it is here that cumulativity (or the lack thereof) can be established. As discussed in §2, a lack of cumulativity in constraint interactions is characterised by a single violation of the highest-ranking

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

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

constraint being prioritised, to the exclusion of all others. The interaction term in the model is one way of measuring this, since it can indicate whether the effect of one violation differs depending on whether it is accompanied by another violation. We expect that, if the effect of one violation is 'cancelled out' in the presence of another, indicating a *lack* of cumulativity, this will be indicated by a significant interaction term with a positive coefficient. If, on the other hand, the main effects of phonotactic violation are significant and the interaction of the two is *not*, we cannot conclude that participants inferred anything but a decrement in log-odds of acceptance or in numerical rating specific to each constraint violated. If, under this scenario, post hoc tests reveal significant differences (in log-odds of endorsement or numerical rating) between each of the singly violating levels and the doubly violating level, this is robust support for cumulative constraint interaction at work. However, if the two main effects of phonotactic violation are significant, and their interaction and the post hoc comparison between singly violating levels and the doubly violating level are not, we cannot conclude that cumulativity was *not* inferred, but neither does the experiment provide strong supporting evidence.

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

5.3 Results

5.3.1 *Binary decision task.* Figure 1 shows the results of the binary decision task in Experiment 1. The final model contained a random intercept for subject and word. There was a main effect of violating backness harmony ($\beta = -0.813$, SE = 0.201, $z = -4.044$, $p < 0.001$) and a main effect of violating nasal harmony ($\beta = -1.748$, SE = 0.202, $z = -8.646$, $p < 0.001$). The interaction between the two was not significant ($\beta = 0.020$, SE = 0.321, $z = 0.061$, $p = 0.951$). This means that, both for backness harmony and for nasal harmony, forms that violated harmony were less likely to be endorsed than those that did not. Further, there is no evidence to think that doubly violating forms were not endorsed at a rate proportional to the summed penalty for each of their violations. Post hoc comparisons indicated that forms violating only nasal harmony trended towards a significant difference from doubly violating forms in log-odds of endorsement ($\beta = 0.833$, SE = 0.474, $z = 1.758$, $p = 0.079$), and that forms violating only backness harmony differed significantly from doubly violating forms ($\beta = 1.768$, SE = 0.475, $z = 3.723$, $p < 0.001$).[7]

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

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

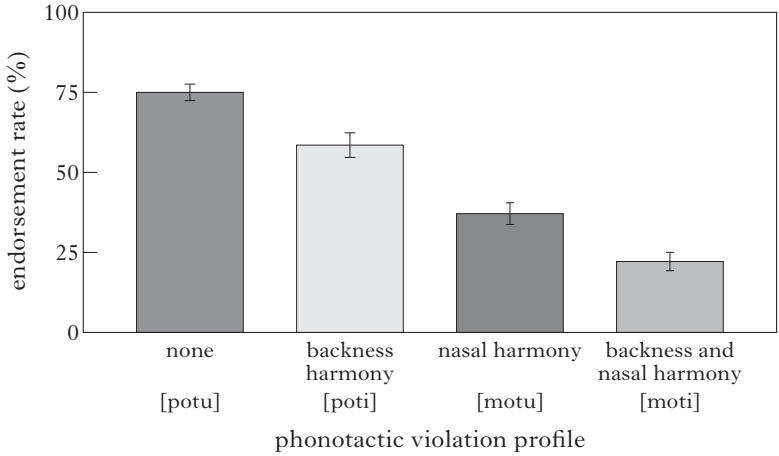

Figure 1

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

5.3.2 *Ratings task*. Results of the ratings task are presented in Fig. 2. The model for the ratings task had the same random effect structure as that of the binary decision task. There was a main effect of violating backness harmony ($\beta = -7.1$, SE = 2.918, $t = -2.433$, $p = 0.020$), which yielded a decrease in ratings. There was also a main effect of violating nasal harmony ($\beta = -23.676$, SE = 2.918, $t = -8.115$, $p < 0.001$); their interaction was not significant ($\beta = -2.025$, SE = 4.614, $t = -0.439$, $p = 0.663$). Post hoc comparisons indicated that forms violating only nasal harmony did not significantly differ in rating from those violating both backness and nasal harmony ($\beta = 5.075$, SE = 6.843, $z = 0.742$, $p < 0.458$), but forms violating only backness harmony did differ from doubly violating forms ($\beta = 21.651$, SE = 6.843, $z = 3.164$, $p = 0.002$).

Experiment 1 provides evidence that, in a sandbox environment, learners infer ganging cumulativity between violations of two separate phonotactic constraints in their learning data. In the binary decision task, doubly violating forms were endorsed in proportion to the likelihood of endorsement of forms having each of their violations independently, while in the ratings task there was a main effect of violating both phonotactics without a significant interaction. Since only backness-violating forms differed from doubly violating forms in the post hoc tests, the evidence on this task is slightly weaker. In either case, however, the overall finding is that cumulativity obtains.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

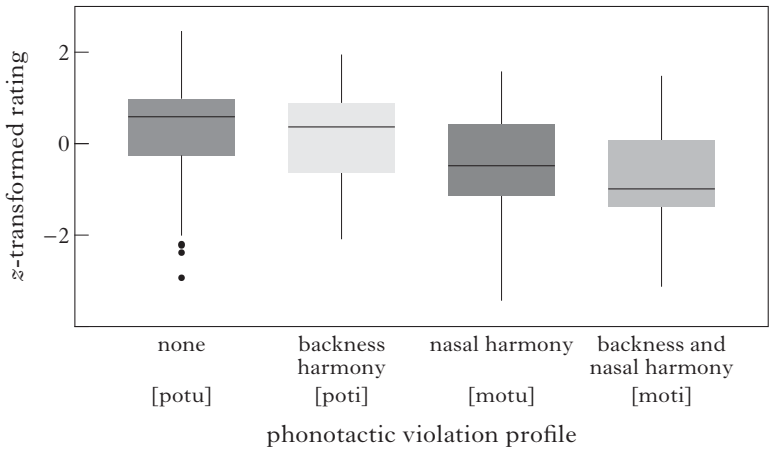

Figure 2

Experiment 1: results for the ratings task. Here and below, the central line in each boxplot indicates the median, with the box extending from the 25th to the 75th percentile; whiskers extend a further 1.5 times the inter-quartile range of the data. For readability, *z*-normalised rating is plotted on the *y*-axis, and the *x*-axis divides the novel words according to their phonotactic violation profile, together with an illustrative example of that profile type.

6 Experiment 2

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 if they had not spoken English since birth ($n = 15$), or did not consistently learn both phonotactic constraints ($n = 35$), leaving 34 participants whose data was included in the study. Note that, although the verification structure for this experiment was identical to that of Experiment 1, this experiment had an extremely high participant-exclusion rate, about 50%. It is an open question as to why this should be – raising the possibility, though by no means the certainty, that the cause could be a substantive difference between sibilant harmony and nasal harmony. This question deserves experimental inquiry beyond the scope of this paper. For present purposes, although the exclusion rate is high, I consider it unlikely that a failure to learn individual phonotactics might affect the way in which these phonotactics – when learned successfully – interact cumulatively in the grammar, and so we

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

can trust the results of the experiment insofar as they are informative about cumulativity.

Recruitment method, compensation, experimental setting and software were the same as for Experiment 1. New materials were created for Experiment 2 by replacing [p] with [ʃ], [m] with [ʒ], [t] with [s] and [n] with [z]. Design, procedure and analysis were identical to that of Experiment 1, except that the binary decision task contained only one presentation of each of the novel words, rather than two, to shorten the experiment and remove the between-block dependency mentioned in note 7.

6.2 Results

6.2.1 *Binary decision task.* Figure 3 shows the results of the binary decision task in Experiment 2. The final logistic regression model contained a random intercept for subject and word. There was a main effect of violating backness harmony ($\beta = -1.223$, SE = 0.177, $z = -6.992$, $p < 0.001$), as well as a main effect of violating sibilant harmony ($\beta = -0.968$, SE = 0.176, $z = -5.497$, $p < 0.001$); the interaction of these factors was not significant ($\beta = 0.283$, SE = 0.281, $z = -1.007$, $p = 0.314$). The post hoc comparison between only backness harmony-violating forms and doubly violating forms was significant ($\beta = 1.506$, SE = 0.415, $z = 3.628$, $p < 0.001$), as was the comparison between only sibilant harmony-violating forms and doubly violating forms ($\beta = 1.251$, SE = 0.414, $z = 3.018$, $p = 0.003$).

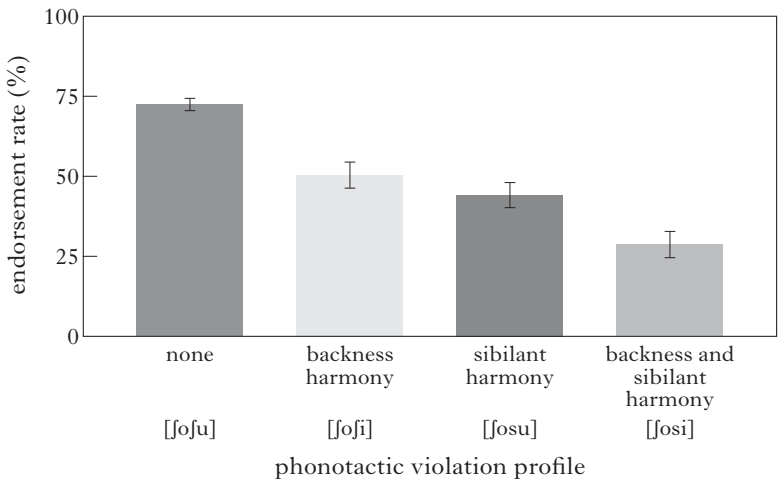

Figure 3
Experiment 2: results for the binary decision task.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

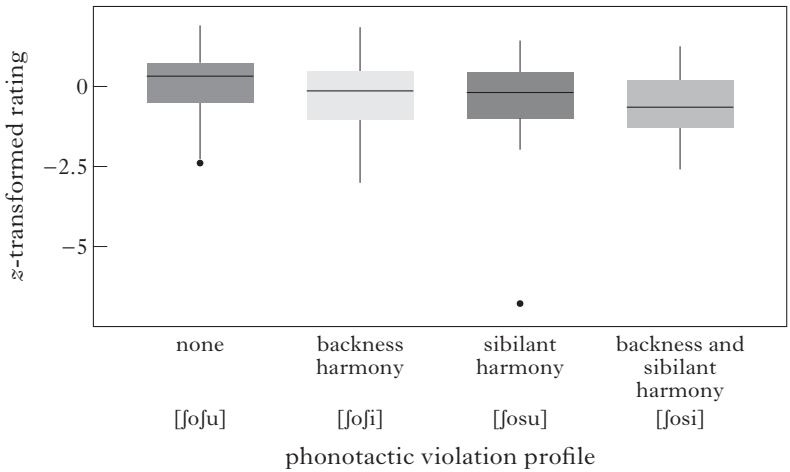

Figure 4
Experiment 2: results for the ratings task.

6.2.2 *Ratings task*. Results of the ratings task are presented in Fig. 4. The final regression modelled raw ratings as a function of violation profile, with random intercepts for subject and word. Mirroring the results from the binary decision task, violation of backness harmony resulted in significantly lower ratings ($\beta = -14.429$, SE = 2.337, $z = -6.070$, $p < 0.001$), as did violations of sibilant harmony ($\beta = -14.722$, SE = 2.374, $t = -6.200$, $p < 0.001$); the interaction of these factors was not significant ($\beta = -1.008$, SE = 3.756, $z = -0.269$, $p = 0.79$). Post hoc comparisons revealed that the difference in rating between only backness harmony-violating forms and doubly violating forms was significant ($\beta = 13.421$, SE = 5.573, $z = 2.408$, $p = 0.016$), as was the difference between only sibilant-violating forms and doubly violating forms ($\beta = 13.713$, SE = 5.569, $z = 2.462$, $p = 0.014$).

The results of Experiment 2 establish the generality of the findings of Experiment 1, confirming that speakers infer ganging cumulativity among different types of phonotactic constraints. Although beyond the scope of this paper, examining a wider range of phonotactics and how they engage in cumulative behaviour would be a valuable contribution to the empirical literature on cumulativity.

7 Experiment 3

Experiment 3 sought to replicate the results of Experiments 1 and 2, using a passive exposure training paradigm designed to more closely mimic first language acquisition. There were two subparts, Experiment 3a and Experiment 3b.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

7.1 Experiment 3a

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

7.1.2 *Design*. The design for Experiment 3a was the same as for Experiment 1, except that the exposure phase consisted of each of the 32 training words, presented in a random order 20 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 if they did not make significantly more errors on verification trials which contrasted in vowel-harmony violation than on those which contrasted in consonant-harmony violation, and *vice versa*. I used Fisher's exact test (Fisher 1934) to determine the level at which the proportion of correct answers for each phonotactic differed significantly, across the range of possible accuracies. The maximum difference between the number of errors participants could make on each type of verification trial without being significantly different by this measure was three. The passive exposure training led to performance on the verification phase which was comparable to that achieved using the more interactive training method (mean accuracy 81.3%), and no subjects were excluded because of a failure to learn both phonotactics to criterion.

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

7.1.3 *Procedure and analysis*. The 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

7.2.1 *Binary decision task*. Figure 5 shows the results of the binary decision task in Experiment 3a. The final logistic regression model contained a random intercept for subject and word. The violation of

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

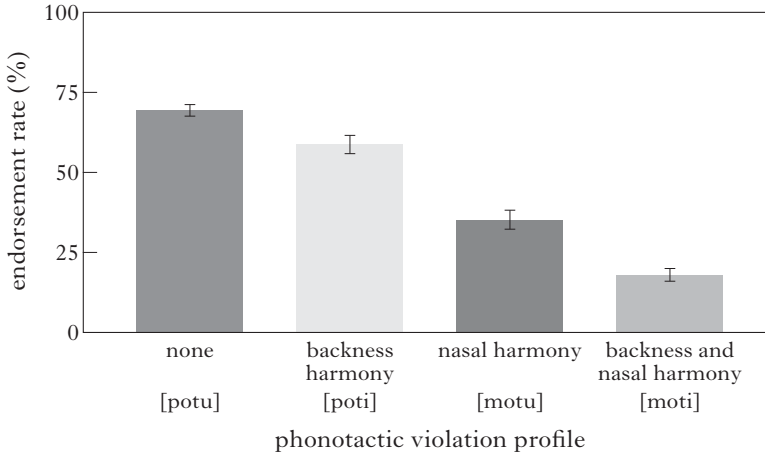

Figure 5
Experiment 3a: results for the binary decision task.

backness harmony was associated with a significant decrease in log-odds of endorsement ($\beta = -0.504$, SE = 0.226, $z = -2.233$, $p = 0.0.026$), as was the violation of nasal harmony ($\beta = -1.562$, SE = 2.227, $z = -6.900$, $p < 0.001$); the interaction between the two was not significant ($\beta = -0.486$, SE = 0.363, $z = -1.292$, $p = 0.196$). Post hoc comparisons revealed that forms that only violated backness harmony were significantly more likely to be endorsed than doubly violating forms ($\beta = 1.094$, SE = 0.533, $z = 2.049$, $p = 0.041$), while forms that violated only nasal harmony were not significantly more likely to be endorsed than doubly violating forms ($\beta = 0.034$, SE = 0.533, $z = 0.065$, $p = 0.948$).

7.2.2 *Ratings task.* Results of the ratings task are presented in Fig. 6. The main effect of violating nasal harmony was significant ($\beta = -17.536$, SE = 3.707, $z = -4.730$, $p < 0.001$), but the main effect of violating backness harmony was not ($\beta = 0.706$, SE = 3.706, $z = 0.190$, $p = 0.850$), nor was the interaction between these factors ($\beta = -8.666$, SE = 5.861, $z = -1.478$, $p = 0.146$). Since the model did not indicate that there was a main effect of violating backness harmony, I did not conduct post hoc tests.

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

7.3 Experiment 3b

In an attempt to better understand the null effect of cumulativity observed in the ratings task in Experiment 3a, a shortened, ratings-only version of

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

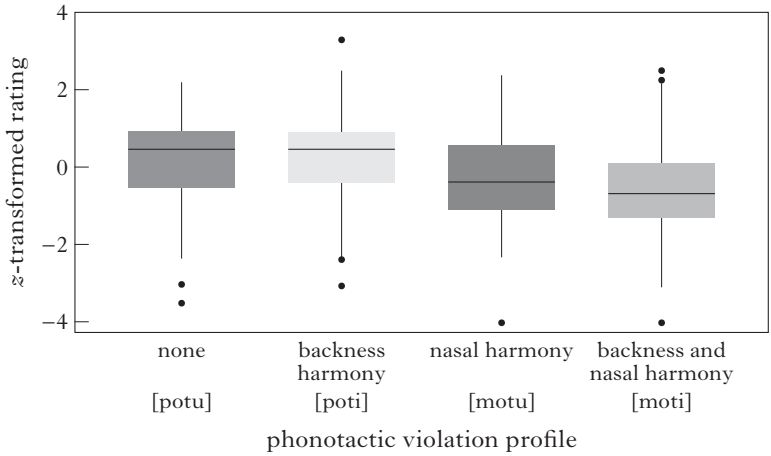

Figure 6
Experiment 3a: results for the ratings task.

the same experiment was carried out. If the results of the ratings task in Experiment 3a were simply the result of random fluctuation in the experimental outcome, we would expect to observe cumulativity in Experiment 3b. If, on the other hand, this difference should be attributed to substantive differences between the designs of the previous three experiments, we would expect to again observe the null result.

7.3.1 *Methods*. 78 undergraduate students were recruited to participate in this experiment, none of whom had participated in the previous experiments. Participants were excluded if they had not spoken English since birth ($n = 7$), did not complete the demographic survey ($n = 1$) or did not consistently learn both phonotactics ($n = 0$), leaving 70 participants whose data was included in the study. Recruitment method, compensation, experimental setting, software, materials and analysis were the same as for Experiment 3a, except that participants did not complete the binary decision task during the test phase.

7.4 Results

The results of the ratings task are presented in Fig. 7. Mirroring the results of the binary decision task from Experiment 3a, there was a main effect of violating backness harmony ($\beta = -10.136$, SE = 3.875, $z = -2.615$, $p = 0.012$) and a main effect of violating nasal harmony ($\beta = -27.029$, SE = 3.875, $z = -6.974$, $p < 0.001$); the interaction between the two was not significant ($\beta = 0.304$, SE = 6.129, $z = -0.050$, $p = 0.961$).

Post hoc comparisons indicated that forms which violated only backness harmony differed significantly in rating from doubly violating forms ($\beta =$

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

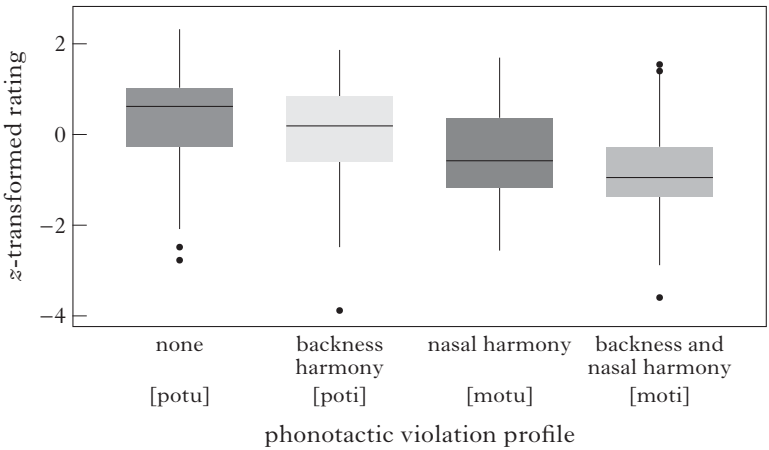

Figure 7
Experiment 3b: results for the ratings task.

27.332, SE = 9.089, $z = -3.007$, $p = 0.003$), and forms which violated only nasal harmony did not ($\beta = 10.439$, SE = 9.089, $z = 1.149$, $p = 0.251$).

Experiment 3b fails to replicate the null effect observed in the ratings task in Experiment 3a – that is, as in Experiments 1 and 2, Experiment 3b showed evidence of ganging cumulativity. I take this to support the hypothesis laid out in §7.3 that the lack of cumulativity observed in Experiment 3a was due to random experimental variation, rather than a substantial difference in the experimental design.

8 Experiment 4

In the previous experiments, 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 but not by OT, counting cumulativity. Because HG takes into account all the violations of each constraint, it predicts that a word violating a constraint n times will be less well-formed than a near-identical word violating the same constraint $n - 1$ times. To test this prediction, participants were taught the exact same two-syllable language used in Experiment 1, but tested on longer novel words, which allowed each word to host up to two violations of each phonotactic constraint. This also allowed me to see whether counting and ganging cumulativity obtained simultaneously, since I examined a number of violations of each single phonotactic in the context of each level of the other, yielding a fully crossed design.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

8.1 Methods

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

8.1.2 *Materials*. Training and verification materials were identical to those of Experiment 1. 48 novel test words were created for this experiment, each four syllables long, with a 'CVCV$_{\text{ı}}$CVCV syllable structure and a left-aligned trochaic stress pattern ([minemeni, putotupo, petipite, …]). 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, for six stimuli per violation-level combination cell.[8]

8.1.3 *Design and procedure*. The design for Experiment 4 was identical to that of Experiment 1. The procedure was identical to that of Experiment 2 (one run through all 48 forms in generalisation per task, rather than two), except that participants were instructed before beginning the test phase that they would be tested on longer words, and that even though the words they would be hearing would be longer than the ones they had learned initially, their length did not bear on whether they were likely to belong to the language or not. This point was stressed via an analogy with English, which contains licit words of many lengths.[9]

8.2 Analysis

The analysis of Experiment 4 differed from that of previous experiments in its design, and because of the additional focus on determining whether learners infer counting cumulativity between constraint violations. Rather than run tests to determine whether each of the nine categories of

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

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

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

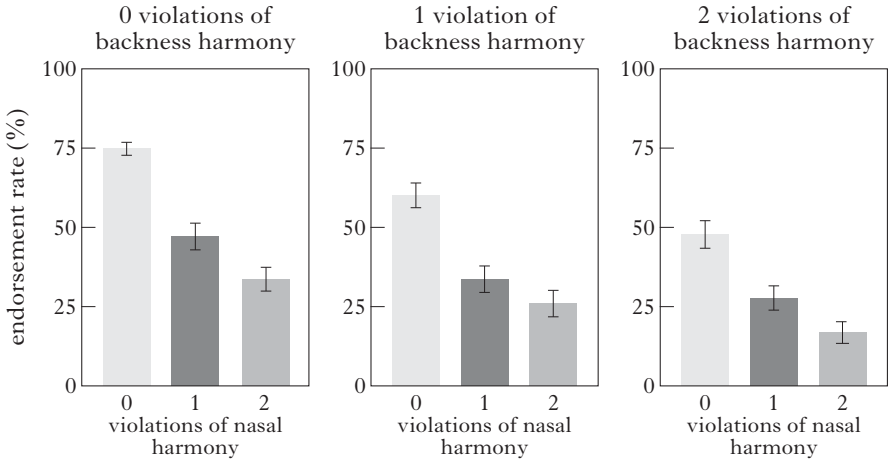

Figure 8

Experiment 4: results for the binary decision task. The *y*-axis plots mean endorsement rate as a percentage, with standard error bars, and the *x*-axis divides the novel words according to their level of vowel-harmony violations, grouping by level of consonant violations.

stimulus in the generalisation phase were significantly different from one another, I fitted a null model, which contained two binary fixed effects (whether or not a form violated each phonotactic) and an alternative model, which contained two three-level factors denoting how many times a form violated each phonotactic (0, 1, 2), and compared these non-nested models using the Akaike Information Criterion (AIC; Burnham & Anderson 2002, 2004). The difference in AIC values between the alternative and null models can be converted into an odds ratio that the alternative model is the one with more explanatory power than the null model. This statistical test directly corresponds to the question that is relevant to linguistic theory: given this data, is a model which allows counting cumulativity more likely than one which does not?

8.3 Results

8.3.1 *Binary decision task.* Figure 8 shows the results of the binary decision task in Experiment 4. The logistic regression models contained a random intercept for subject and word. Two versions of this model were fitted, the NULL model, with a binary factor (Violating *vs.* Non-violating), and an ALTERNATIVE model, with a three-level factor corresponding to violation level (0, 1, 2) of each phonotactic, discussed above. The only difference between these two models is that the alternative model allows for a distinction between multiple levels of violation – counting cumulativity – and the null model does not. The odds of the alternative

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

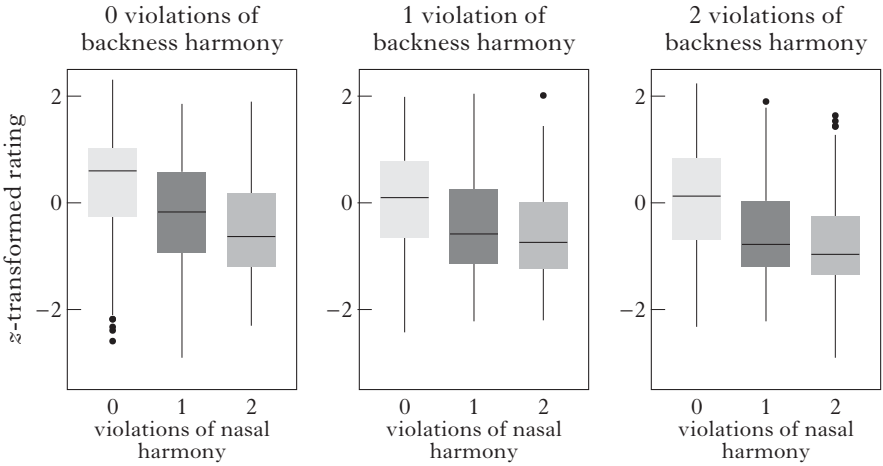

Figure 9
Experiment 4: results for the ratings task.

model being superior are ≈141:1 (ΔAIC = 9.9). I take this as evidence in favour of counting cumulativity playing a role in generating the experimental data.

8.3.2 *Ratings task.* Figure 9 shows the results of the ratings task in Experiment 4. Null and alternative models of the same structure as those described above were fitted to the ratings data; AIC-based model comparison indicated that the odds of the alternative model being superior were ≈33:1 (ΔAIC = 7), again providing support for cumulativity.

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

9 Discussion and conclusion

This paper used a series of artificial grammar learning experiments to investigate how learners acquire multiple phonotactic generalisations simultaneously, and how these generalisations 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.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

Experiment 2 replicated these findings with 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 generalise their knowledge to longer words, and demonstrated that participants infer counting cumulativity as well. The potential significance of these results is that they demonstrate cumulative effects on phonotactic well-formedness that cannot be explained by lexical frequency asymmetries. Further, they demonstrate that in the absence of evidence for or against constraint interaction, learners behave as expected if they have grammars in which constraint cumulativity is the norm, and in ways which are explicitly predicted not to be possible by grammars incapable of expressing cumulative relationships.

9.1 Implications for phonological frameworks

As discussed in §2, phonological frameworks differ in their generative capacity to capture cumulative constraint interactions. By design, classical OT and other strict-domination-based frameworks rule out both counting and ganging cumulativity. On exception to this rule is Stochastic OT (Boersma 1997, Boersma & Hayes 2001), which, while able to capture ganging cumulativity within a certain range (cf. Zuraw & Hayes 2017, Kawahara 2020, Smith & Pater 2020), is unable to capture counting cumulativity for the same reason that classical OT is unable to – a single constraint is either violated or satisfied, and no notion of 'number of times violated' exists beyond what is needed to determine which constraint is violated *more* often when no candidates are violation-free (Prince & Smolensky 1993: 18). While in-depth model comparison is not carried out here, this data suggests the tentative conclusion that only weighted-constraint frameworks such as Maximum Entropy Harmonic Grammar (Smolensky 1986, Goldwater & Johnson 2003) and Noisy Harmonic Grammar (Boersma & Pater 2016) are adequately expressive models of the phonological grammar.

9.2 Future work

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

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

REFERENCES

Albright, Adam (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology* **26**. 9–41.
Albright, Adam (2012). Additive markedness interactions in phonology. Ms, MIT.
Bailey, Todd M. & Ulrike Hahn (2001). Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language* **44**. 568–591.
Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tilly (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* **68**. 255–278.
Bates, Douglas, Martin Mächler, Benjamin M. Bolker & Steven C. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**. 1–48.
Becker, Michael & Jonathan Levine (2020). Experigen: an online experiment platform. Available (November 2020) at https://github.com/tlozoot/experigen.
Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**. 43–58.
Boersma, Paul & Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *LI* **32**. 45–86.
Boersma, Paul & Joe Pater (2016). Convergence properties of a Gradual Learning Algorithm for Harmonic Grammar. In John J. McCarthy & Joe Pater (eds.) *Harmonic Grammar and Harmonic Serialism*. London: Equinox. 389–434.
Burnham, Kenneth P. & David R. Anderson (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. 2nd edn. New York: Springer.
Burnham, Kenneth P. & David R. Anderson (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research* **33**. 261–304.
Coetzee, Andries W. & Joe Pater (2006). Lexically ranked OCP-Place constraints in Muna. Ms, University of Michigan & University of Massachusetts, Amherst. Available as ROA-842 from the Rutgers Optimality Archive.
Coleman, John & Janet B. Pierrehumbert (1997). Stochastic phonological grammars and acceptability. In John Coleman (ed.) *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*. Somerset, N.J.: Association for Computational Linguistics. 49–56.
Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann (2011). Explaining sonority projection effects. *Phonology* **28**. 197–234.
Durvasula, Karthik & Adam Liter (2020). There is a simplicity bias when generalizing from ambiguous data. *Phonology* **37**. 177–213.
Featherston, Sam (2005). Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua* **115**. 1525–1550.
Featherston, Sam (2019). The decathlon model. In András Kertész, Edith Moravcsik & Csilla Rákosi (eds.) *Current approaches to syntax: a comparative handbook*. Berlin & New York: De Gruyter Mouton 155–185.
Finley, Sara (2015). Learning nonadjacent dependencies in phonology: transparent vowels in vowel harmony. *Lg* **91**. 48–72.
Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London* (*Series A*) **144**. 285–307.
Frisch, Stefan A., Nathan R. Large & David B. Pisoni (2000). Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* **42**. 481–496.
Fukazawa, Haruka, Shigeto Kawahara, Mafuyu Kitahara & Shinichiro Sano (2015). Two is too much: geminate devoicing in Japanese. *On-in Kenkyu* **18**. 3–10.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

Gallagher, Gillian (2013). Learning the identity effect as an artificial language: bias and generalisation. *Phonology* **30**. 253–295.
Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. 111–120.
Guy, Gregory R. (1997). Violable is variable: optimality theory and linguistic variation. *Language Variation and Change* **9**. 333–347.
Hansson, Gunnar Ólafur (2010). *Consonant harmony: long-distance interaction in phonology*. Berkeley: University of California Press.
Hothorn, Torsten, Frank Bretz, Peter Westfall, Richard M. Heiberger, Andre Schuetzenmeister & Susan Scheibe (2016). Multcomp: simultaneous inference in general parametric models. R package version 1.4-15. http://cran.r-project.org/web/packages/multcomp/index.html.
Itô, Junko & Armin Mester (1986). The phonology of voicing in Japanese: theoretical consequences for morphological accessibility. *LI* **17**. 49–73.
Jäger, Gerhard & Anette Rosenbach (2006). The winner takes it all – almost: cumulativity in grammatical variation. *Linguistics* **44**. 937–971.
Jarosz, Gaja & Amanda Rysling (2017). Sonority sequencing in Polish: the combined roles of prior bias & experience. In Karen Jesney, Charlie O'Hara, Caitlin Smith & Rachel Walker (eds.) *Proceedings of the 2016 Meeting on Phonology*. http://dx.doi.org/10.3765/amp.v4i0.3975.
Kawahara, Shigeto (2011a). Aspects of Japanese loanword devoicing. *Journal of East Asian Linguistics* **20**. 169–194.
Kawahara, Shigeto (2011b). Japanese loanword devoicing revisited: a rating study. *NLLT* **29**. 705–723.
Kawahara, Shigeto (2012). Lyman's Law is active in loanwords and nonce words: evidence from naturalness judgment studies. *Lingua* **122**. 1193–1206.
Kawahara, Shigeto (2013). Testing Japanese loanword devoicing: addressing task effects. *Linguistics* **51**. 1271–1299.
Kawahara, Shigeto (2020). A wug-shaped curve in sound symbolism: the case of Japanese Pokémon names. *Phonology* **37**. 383–418.
Kawahara, Shigeto & Canaan Breiss (to appear). Exploring the nature of cumulativity in sound symbolism: experimental studies of Pokémonastics with English speakers. *Laboratory Phonology* **12**.
Kawahara, Shigeto & Shin-ichiro Sano (2016). /p/-driven geminate devoicing in Japanese: corpus and experimental evidence. *Journal of Japanese Linguistics* **32**. 53–73.
Kim, Seoyoung (2019). Modeling self super-gang effects in MaxEnt: a case study on Japanese Rendaku. *NELS* **49:2**. 175–188.
Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* **82**. http://dx.doi.org/10.18637/jss.v082.i13.
Lai, Regine (2015). Learnable *vs*. unlearnable harmony patterns. *LI* **46**. 425–451.
Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990). Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Hillsdale: Erlbaum. 388–395.
Martin, Andrew (2007). *The evolving lexicon*. PhD dissertation, University of California, Los Angeles.
Martin, Andrew (2011). Grammars leak: modeling how phonotactic generalizations interact within the grammar. *Lg* **87**. 751–770.
Nishimura, Kohei (2003). *Lyman's Law in loanwords*. MA thesis, Nagoya University.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press

Pater, Joe (2009). Weighted constraints in generative linguistics. *Cognitive Science* **33**. 999–1035.
Pierrehumbert, Janet B. (to appear). 70+ years of probabilistic phonology. In B. Elan Dresher & Harry van der Hulst (eds.) *The Oxford handbook of the history of phonology*. Oxford: Oxford University Press. Preliminary version available (December 2020) at http://www.phon.ox.ac.uk/jpierrehumbert/publications.html.
Pizzo, Presley (2015). *Investigating properties of phonotactic knowledge through web-based experimentation*. PhD dissertation, University of Massachusetts Amherst.
Prince, Alan & Paul Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Ms, Rutgers University & University of Colorado, Boulder. Published 2004, Malden, Mass. & Oxford: Blackwell.
R Core Team (2020). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.
Rose, Sharon & Lisa King (2007). Speech error elicitation and co-occurrence restrictions in two Ethiopian Semitic languages. *Language and Speech* **50**. 451–504.
Shademan, Shabnam (2007). *Grammar and analogy in phonotactic well-formedness judgments*. PhD thesis, University of California, Los Angeles.
Shih, Stephanie S. (2017). Constraint conjunction in weighted probabilistic grammar. *Phonology* **34**. 243–268.
Smith, Brian W. & Joe Pater (2020). French schwa and gradient cumulativity. *Glossa* **5**(1):24. http://doi.org/10.5334/gjgl.583.
Smolensky, Paul (1986). Information processing in dynamical systems: foundations of Harmony Theory. In D. E. Rumelhart, J. L. McClelland & the PDP Research Group (eds.) *Parallel Distributed Processing: explorations in the micro-structure of cognition*. Vol. 1: *Foundations*. Cambridge, Mass.: MIT Press. 194–281.
Smolensky, Paul (1993). Harmony, markedness, and phonological activity. Paper presented at Rutgers Optimality Workshop. Available as ROA-87 from the Rutgers Optimality Archive.
Smolensky, Paul & Géraldine Legendre (eds.) (2006). *The harmonic mind: from neural computation to optimality-theoretic grammar*. Vol. 1. Cambridge, Mass.: MIT Press.
Walker, Rachel (2011). *Vowel patterns in language*. Cambridge: Cambridge University Press.
Wilson, Colin & Gillian Gallagher (2018). Accidental gaps and surface-based phonotactic learning: a case study of South Bolivian Quechua. *LI* **49**. 610–623.
Zuraw, Kie & Bruce Hayes (2017). Intersecting constraint families: an argument for Harmonic Grammar. *Lg* **93**. 497–548.

https://doi.org/10.1017/S0952675720000275 Published online by Cambridge University Press