Cumulative markedness effects and (non-)linearity in phonotactics*

Canaan Breiss[a], Adam Albright[b]

[a]*University of California, Los Angeles*
[b]*Massachusetts Institute of Technology*

Abstract

This study uses an Artificial Grammar Learning experiment to test for a synchronic relationship between the severity of an individual phonotactic violation and the linearity of its cumulative interaction with other violations, prompted by previous experimental findings (Albright, 2012, Breiss, *submitted*). We find that as individual phonotactic patterns are made more exceptionful, their interaction moves from linear to super-linear, and argue that this provides evidence for a non-linear relationship between Harmony and probability. We evaluate five contemporary phonological frameworks using this data, and find that those which incorporate such a non-linear relationship – Maximum Entropy HG and Noisy HG – are able to capture the super-linear patterns observed significantly better than other frameworks. Further, we demonstrate that a MaxEnt model provided the same training data as experimental participants exhibits similar emergent super-linear cumulativity, and explore the weighting conditions under which MaxEnt models yield sub-linear, linear, and super-linear cumulativity.

Keywords: phonotactics, super-linear, super-additive, cumulative constraint interaction, artificial grammar learning, gang effect

*Thanks to Bruce Hayes, Gaja Jarosz, Giorgio Magri, Kie Zuraw, and members of audiences at the 2020 LSA Annual Meeting and the UCLA Phonology Seminar for helpful discussion and feedback. Thanks also to Beth Sturman for recording stimuli, and to Ruby Dennis, Gabrielle Dinh, 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 authors. This work was supported by NSF Graduate Research Fellowship DGE-1650604 to Canaan Breiss.

1. Introduction

This paper addresses the relationship between the strength of phonotactic restrictions and the way in which multiple coincident violations of such restrictions interact in the grammar. Prompted by a disconnect between previous experimental results, we investigate whether there is a causal relationship between the strength of a given phonotactic restriction (as measured in number of exceptions) and how its penalty combines with that of other violations in the grammar. Using an Artificial Grammar Learning (AGL) paradigm similar to Breiss (*submitted*), we find that as the number of exceptions to a given phonotactic increases, the additional penalty for multiple coincident violations increases beyond what is obtained by simple addition of these independent penalties. That is, participants' acceptability ratings for doubly-marked forms are *lower* than expected based on the independent penalties for each of these forms' individual violations. We argue that this constitutes a case of *super-linear cumulativity*, implying a non-linear relationship between the psychological quantity Harmony as computed in the grammar, and experimentally-assessed acceptability/probability.

This finding in hand, we consider what characteristics a phonological framework must possess to capture these effects, and use the experimental data to compare two prominent constraint-based models, Maximum Entropy Harmonic Grammar (MaxEnt; Smolensky (1986); Goldwater and Johnson (2003)) and Noisy Harmonic Grammar (NHG; Boersma and Pater (2008)), which exhibit such a non-linear relationship between Harmony and probability. We also evaluate a range of other constraint-based theories of phonotactic well-formedness, Stochastic OT (Boersma et al., 1997, 1998; Boersma and Hayes, 2001), Linear Harmonic Grammar (Coetzee and Pater (2008), cf. also the nearly-identical Linear Optimality Theory of Keller (2006)), and the surface-based Maximum Entropy approach (hereafter simply called the Surface-based model (Hayes and Wilson, 2008; Wilson and Gallagher, 2018)). We find that MaxEnt and NHG can capture the non-linear relationship substantially better than the other frameworks, and further that MaxEnt outperforms NHG. Finally, we demonstrate that the MaxEnt model can arrive at a weighting which displays super-linear cumulativity when provided only with the training data participants received.

2. How do speakers compute grammaticality across multiple marked structures?

A growing consensus in the phonological literature supports the view that markedness violations are *cumulative*: when speakers judge the well-formedness of a word, their judgement is not based solely on the most marked structure it contains (as predicted by strict-ranking constraint-based models such as Optimality Theory (Prince and Smolensky, 1993) and its variants). Rather, they attend to all relevant structures and weight their importance according to their severity (as predicted by weighted-constraint models such as Harmonic Grammar (Legendre et al., 1990) and its variants). This aggregation of evidence — termed *cumulativity* (cf. Jäger and Rosenbach (2006) for more on this terminology) — can be observed at two levels: that of probability of a given structure in the lexicon, and that of experimentally-determined acceptability or probability. Lexically, it has been repeatedly observed that the frequency of word-types which contain a single given marked structure x is greater than that of words which contain both x and also an independently marked structure y. An example of this type of cumulativity can be found in the lexicon of English: as part of a study of English monosyllable phonotactics, Albright (2012) found that 491 (8.2%) of monosyllables in the CELEX database (Baayen et al., 1995) begin with a *stop+l* sequence, and 47 (3.2%) end with a *s+stop* sequence. However, the number of *#stop+l…s+stop#* words is lower than either of these, with only 7 occurrences (0.11%). Experimentally, Pizzo (2015) found that words which violated English syllable-margin phonotactics in one location, ex. *pla<u>vb</u>, <u>tl</u>ag* were judged less acceptable than one which violated none — *plag* — and crucially were rated *more* acceptable than those which violated both, ex., *<u>tl</u>a<u>vb</u>*. Breiss (*submitted*) tested for cumulativity in phonotactic markedness using an Artificial Grammar Learning paradigm, and found that when trained on a language which conformed to two exceptionless phonotactics participants judged words that violated both phonotactics less well-formed than those which violated only one, again demonstrating cumulativity.

2.1. A note on terminology

This paper builds on the repeated experimental observations of cumulativity to probe *how* exactly independent markedness violations combine their effects in the grammar. Thus, it is important to clearly define terms describing different types of cumulativity.

In general, discussions of constraint cumulativity in Harmonic Grammar have assumed that when a word violates multiple constraints, the Harmony of that word is obtained by summing the weighted violations of all constraints. This assumption that each markedness violation makes independent discrete contributions to Harmony is followed here as well. Theories differ, however, on how Harmony is related to probability or acceptability (used here interchangeably in experimental contexts, in keeping with the linking hypothesis adopted in section 7.1.1, although in general this distinction is non-trivial). Focusing on probabilistic frameworks, some theories such as Linear Harmonic Grammar (Linear HG) hold that the probability of a form is proportional to its Harmony. This is what we term *linear cumulativity*: a change in Harmony translates into a proportional change in probability regardless of how marked the form was to start with. Other theories, such as MaxEnt and NHG, hold that there is a sigmoidal, non-linear relationship between Harmony and probability: under certain weighting conditions, these models yield *super-linear cumulativity* — the same change in Harmony translates into a larger change in probability when it occurs in the context of other constraint violations that when it occurs alone. For the same mathematical reason, these models also predict *sub-linear cumulativity* — the same change in harmony translates into a smaller change in probability when it coincides with other violations than when it occurs alone.[1] Though the majority of this paper focuses on super-linear behavior, we return to the topic of sub-linearity in section 8.2. Finally, since this paper focuses on evidence which can distinguish between linear and super-linear accumulation of penalties in the dependent variable, we must define a method for computing the probability of a form under *linear cumulativity* of markedness, so as to have a baseline to compare to. We do this by taking the sum of the decrements in probability associated with each of the independent Harmony decrements that each constraint violation the form incurs. This is equivalent to saying that under linear cumulativity, the probability of a form is equal to the joint probability of the structures it contains: $p(\mathit{form}|\text{Violates Constraint A and B}) = p(\mathit{form}|\text{Violates Constraint A}) \times p(\mathit{form}|\text{Violates Constraint B})$.[2]

3. Super-linear cumulativity in experimental data, and how to explain it

Given these definitions, what is the empirical landscape of cumulativity in the phonological literature? Recent work by Breiss (*submitted*) found using an AGL

[1]We note that while in general ranked-constraint theories do not allow cumulative effects, Classical Optimality Theory's strict-domination can be understood as an extreme case of sub-linear cumulativity of markedness penalties where only the penalty associated with violating the highest-ranked constraint distinguishes candidates.

[2]In the case that some continuous measure of acceptability is the dependent variable, we can sum the independent decrements in the continuous measure (assuming the scale it is measured on is itself linear in the conventional sense) to find the expected decrement for those coincident violations.

paradigm that participants inferred linear cumulativity between markedness violations. However, this finding of linearity contrasts with a number of cases of apparent super-linear cumulativity in other lexical and experimental data. For an illustrative example, let us return to Albright's study of monosyllabic English words. In section 1 the example served to illustrate the presence of *any kind* of cumulativity in lexical counts — that words with two marked structures were less common than words with only one. Quantifying what "less common" means, however, reveals that in this instance the cumulativity exhibited is *super-linear* in nature: the independent probabilities of the marked syllable margins alone predicts that $8.2\% \times 3.2\% = 0.22\%$ of the monosyllables in the database — about 16 unique words — should exhibit both the marked onset and marked coda. In fact, however, there are only 7 words which do so, less than half the number predicted by the joint probability of the marked structures.

To complement these lexical findings, Albright (2012) replicated a nonword acceptability judgment task from Bailey and Hahn (2001) which asked subjects to rate the acceptability of novel English monosyllables containing onset clusters (e.g. [krɛn, draf]), coda clusters (e.g. [lɛsk, mɪsp]), or both (e.g. [drɪsp, krɛsk]). Albright then modeled whether the acceptability of the doubly-marked forms could be predicted solely on the basis of their constituent violations — the expected result if Harmony is linearly related to acceptability — and found that it could not: doubly-marked forms such as [drɪsp] were rated *less acceptable* than predicted by the sum of their independent penalties. Other cases of super-linearity have been documented in phonological alternations: for example Smith and Pater (2020) note that super-linear behavior is observed in the interaction of deletion and epenthesis in the surface-realization of French schwa.

3.1. In search of an explanatory theory

Although there is growing evidence for super-linear cumulativity in experimental studies, there is not yet a consensus about the synchronic mechanism that produces these effects. A prerequisite for any suitable explanation is that it must be able to explain simultaneously both established examples of linear cumulativity (findings of Breiss (*submitted*), and literature reviewed therein) as well as the super-linear cumulativity reviewed here. Below, we outline two theoretical mechanisms which yield testable predictions about how and when we should expect super-linear effects (or lack thereof) in constraint interactions.

3.1.1. *The lexicon as an experimental confound*

Although the studies discussed in section 3 are suggestive of a synchronic non-linear relationship between Harmony and acceptability, the close connection between the lexicon and the phonological grammar prevents us from drawing reliable conclusions. A robust body of evidence, primarily drawn from the study of phonological alternations, indicates that all else equal, speakers will extend statistical generalizations which hold of their lexicon to novel items in generalization tasks such as *wug* tests, a propensity which Hayes et al. (2009) call the "Law of Frequency Matching". This tendency poses a confound for the study of super-linear cumulativity, however, in light of literature which suggests that the lexicons of many languages exhibit super-linear under-representation (cf. Albright (2012)'s own findings on English; a larger review of similar lexical findings is postponed to section 8.3). Thus there is an alternative explanation for experimental findings such as those of Albright (2012), namely that they could be an experimental artefact of participants generating acceptability judgements using phonological grammars learned from a lexicon that exhibits super-linear underrepresentation of the very structures which the participants are being asked to judge. In this scenario, the synchronic grammar would not need to encode a relationship between exceptionality and super-linearity, only to allow participants to draw on fine-grained statistical information about the lexicon, a capacity which has been well established over the past decade or more of phonological research. In this scenario, the experimental findings of Albright and others in section 3 would be due to frequency-matching lexical statistics which, for reasons independent of the synchronic grammar, exhibit such a relationship, while the linear cumulativity for exceptionless phonotactics found by Breiss would reflect the synchronic phonological grammar.

3.1.2. *A synchronic mechanism for super-linear cumulativity*

In contrast to the lexical account, certain phonological frameworks predict that super-linear cumulativity should emerge from the synchronic grammar under certain constraint weighting conditions. Specifically, as noted in section 2.1, MaxEnt and NHG models predict a nonlinear, sigmoidal relationship between Harmony and probability; for in-depth discussion, see Smith and Pater (2020); Zuraw and Hayes (2017); Hayes (2017, 2020); McPherson and Hayes (2016). This phenomenon is illustrated with the schematic diagram in Figure 1.

The horizontal axis counts violations of a generic scalar markedness constraint, and the vertical axis plots the probability of a candidate violating it competing against a candidate with a single violation of a binary faithfulness con-

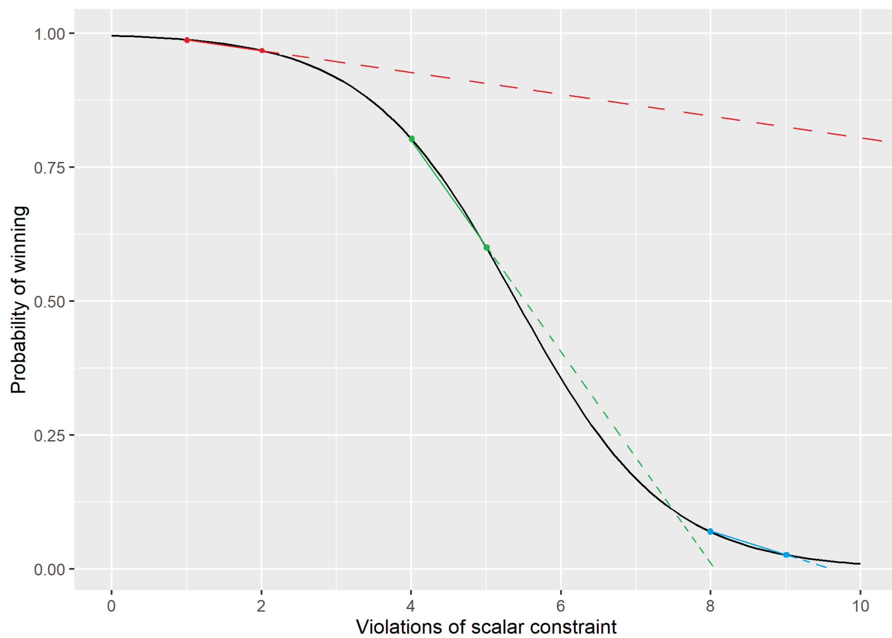

Figure 1: A schematic illustration of a sigmoidal relationship between Harmony (horizontal axis) and Probability (vertical axis). The red, green, and blue lines are tangent to the curve, and demonstrate super-linear (red), linear (green), and sub-linear (blue) cumulativity.

straint (this schematization is heavily inspired by Zuraw and Hayes (2017); Hayes (2020)). As the number of scalar constraint violations increases, the probability of the candidate violating the scalar markedness constraint decreases in a non-linear fashion. In this scenario, the relationship between different numbers of violations of the scalar markedness constraint are non-linear, technically reminiscent of *counting cumulativity* (cf. Jäger and Rosenbach (2006) on this terminology). To see how the same mechanism applies in cases of *ganging cumulativity*, we need only consider the case where the horizontal axis plots the Harmony of a candidate which is a composite of contributions from several different markedness violations. Here the non-linearity of the relationship between Harmony and probability is evident in that the same amount of Harmony penalty yields differently-sized decrements in probability at different places on the horizontal axis. This corresponds to the different slopes of the tangent lines in figure 1.

The proposed synchronic mechanism can account for both the super-linear cumulativity of exceptionful phonotactics found by Albright as well as the linear cumulativity of those found by Breiss, because it directly encodes a relationship between the strength of a phonotactic and the nature of the cumulativity that it enters into. We return to the specific mathematical characterization of this relationship in in MaxEnt in section 8.1.

4. Experimental design overview

To distinguish between these two accounts, we used an AGL experiment to test the prediction about the relationship between constraint weight and linearity made by the MaxEnt and NHG frameworks. If the experimentally-observed super-linear cumulativity discussed above has its source in the synchronic grammar, the linearity of constraint interaction should be sensitive to the number of exceptions to the phonotactics involved (as a proxy for constraint "strength"). If the synchronic grammar plays no role, however, such a manipulation should not impact the cumulativity that the constraints exhibit.

Breiss (*submitted*) used a poverty-of-the-stimulus paradigm to teach participants a language containing evidence for two phonotactic constraints, who were then asked to give well-formedness ratings of novel words violating neither, one, or both phonotactics. In Experiment 1 we adopt a very similar experimental design, except that we systematically manipulate the number of exceptions to each phonotactic constraint. The critical manipulation was the training group to which participants were assigned: we created five distinct training Conditions (labeled A-E for convenience), each containing a different number of exceptions to each phonotactic (ranging from 0% (Condition A) to 25% (Condition E) in 6.25% increments). The quantity of interest was the penalty for doubly-marked forms relative to the singly-marked ones, as modulated by the percentage of learning data which did not conform to the majority pattern.

An advantage of pursuing this question using an AGL task rather than studying speakers' intuitions about their native language, as was done in Albright (2012), is that the total number of items in the "lexicon" participants are exposed to is very small. This makes it easier to maintain the poverty-of-the-stimulus environment that would otherwise be compromised by speakers' knowledge of their lexicon, either because they could infer a specific number of doubly-marked structures based on the joint probability of those minority structures in the lexicon, or have noticed an obvious gap in the lexicon corresponding to doubly-marked forms and learned a more complex phonotactic penalizing them directly. Within the tight confines of the artificial lexicon, however, we can keep the number of doubly-marked forms expected via the joint probability of the individual marked structures very close to zero (in the experiment reported below it ranges between 0 and 2 in 0.25-unit increments across five conditions). Thus the experiment is ambiguous as to whether the nonoccurence of doubly-marked forms is due to a linear or super-linear cumulativity. Experiment 2 further addresses the question of whether participants could have been sensitive to the fraction of expected doubly-marked forms absent from the more exceptionful conditions of Experiment 1.

5. Experiment 1: testing for a synchronic mechanism of super-linear cumulativity

This experiment built upon the results of an experiment reported as Experiment 3b in Breiss (*submitted*); data from that experiment is included here as Condition A, the training group whose learning data conformed exceptionlessly to both phonotactics. All data reported below included these participants. Further, the experimental materials were the same as those used in Breiss (*submitted*) unless reported otherwise.

5.1. Methods

5.1.1. Participants

375 undergraduate students were recruited from the Psychology Subject Pool at a North American university, and were compensated with course credit. Participants' data were excluded if they failed to meet the criterion for sufficient learning as assessed during the verification phase ($n = 0$; see section 5.2 for details), for not having spoken English consistently since early childhood ($n = 43$), and in the case of experimenter error ($n = 3$), leaving data from 329 participants included in the final analysis.

5.1.2. Stimuli

The exposure phase contained 32 CVCV, initially-stressed nonwords, with consonants /p, t, m, n/ and vowels /i, e, u, o/. One of the two phonotactics was a requirement that consonants harmonize with respect to the feature [nasal], such that both consonants in the word were drawn from either /p, t/ or /m, n/ (exhibiting *nasal harmony*). The other phonotactic required that vowels harmonize with respect to the feature [back], such that both vowels in the word were drawn from either /i, e/ or /u, o/ (*backness harmony*). For more on these types of consonant and vowel harmony respectively, see Hansson (2010); Walker (2011).

Five distinct training Conditions (A-E) were distinguished by the number of items in training that violated each of the phonotactic patterns in the language:

Condition:		A	B	C	D	E
Percent exceptions:		0%	12.5%	25%	37.5%	50%
Unmarked	*potu*	32	28	24	20	16
Back exceptions	*poti*	0	2	4	6	8
Nasal exceptions	*ponu*	0	2	4	6	8
Doubly-violating	*poni*	0	0	0	0	0

Table 1: Distribution of stimuli across Conditions in Experiment 1.

0%, 6.25%, 12.5%, 18.75% or 25%. There were no training items which violated both phonotactics at once, so even in the most exceptionful Condition (Condition E) each phonotactic received support from 75% of the words in the training phase. Table 1 below displays the counts and violation profiles of stimuli.

The verification phase used 16 pairs of minimally-differing nonwords: one member of each pair was a fully-conforming word from the exposure phase, and the other was created by reversing the featural specification for backness or nasality of one of the consonants or vowels in the fully-conforming word. This yielded a pair of words differing only in a single instance of that feature. 8 pairs differed in a violation of nasal harmony, and 8 in violation of backness harmony, with differences between pair-members balanced for segmental placement and identity. Verification pairs were balanced so that when a fully-conforming verification word had identical consonants (ex. *totu*), it differed differed only in the violation of backness harmony (ex., *totu* vs. *toti*). The same condition was imposed on verification trials whose conforming word contained identical vowels. There were no doubly-violating words in the verification phase, since its purpose was simply to ensure that participants had learned each of the two phonotactic constraints independently.

The test phase used a set of 48 novel nonwords which varied in conformity both phonotactics. 24 conformed to both phonotactics (ex. *potu*), eight violated only the nasal-harmony phonotactic (ex., *ponu*), eight violated only the backness-harmony phonotactic (*poti*), and eight violated both the nasal-harmony and backness-harmony phonotactics (*poni*).

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

5.2. Design

Participants were assigned to one of the five Conditions, and learned the language by listening to a continuous speech stream containing 20 randomized repetitions of the 32 words selected for that particular training phase. After exposure, participants completed 16 self-paced two-alternative forced choice verification trials. Participants were allowed to advance to the generalization phase if they learned each of the phonotactics to the same degree. This was operationalized by imposing a condition that the difference in number of correct answers between pairs differing only in a nasal harmony violation and those differing only in a backness harmony violation was not allowed to be greater than 2, chosen by using 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. If participants did not meet criteria after two exposure blocks (one initial and one after failing to meet criterion during the verification phase), they were simply asked to complete the final demographic questionnaire and did not generate data in the generalization phase (although recall from in section 5.1.1 that no participants were excluded for this reason).

If participants met criteria on the verification phase, they advanced to a generalization phase which consisted of a ratings task containing 48 novel words in which participants were asked to rate each of the words on a scale from 0 (*very bad*) to 100 (*very good*) based on how good they sounded as an example of the language they had learned during the exposure phase. At the end of the experiment, demographic and language-background information was collected. The entire experiment lasted approximately 20-30 minutes, depending on the number of additional exposure blocks each participants required.

5.2.1. Procedure

The experiment was conducted in a sound-attenuated room using a modified version of the Experigen platform (Becker and Levine, 2010). At the start of the experiment, participants were informed that they would first be learning a new language, and that they then would be tested on their knowledge of that language. During the exposure phase, participants were instructed to simply sit and listen to the speech stream and, if they felt themselves getting bored, to try to count how many unique words they could find in the speech stream (this task was suggested simply to encourage participants to attend to task). The exposure phase lasted about ten minutes.

Following the exposure phase, participants completed a self-paced verification phase. On each verification trial participants were played a pair of nonwords in a random order, and were instructed to choose the one that sounded like it could belong to the language they had learned. The generalization phase followed a similar structure, except that each trial containing a single novel nonword to which participants assigned a numerical rating. After completing the generalization phase (or after failure to meet criterion during the verification phase), participants completed a brief demographic questionnaire designed to assess their language background.

5.3. Analysis

To determine the effect of the number of nonconforming items in training effect on the linearity of the constraint interaction, we fit a linear mixed effects regression model using the `lme4` package (Bates et al., 2015a) in R (Team et al., 2013) to the ratings data from the generalization phase. The model included a three-way interaction between two binary fixed effects (violation of vowel harmony (*y*/*n*), violation of consonant harmony (*y*/*n*)) and a continuous fixed effect corresponding to the percentage of exceptions to individual phonotactics in a given participants' training condition (hereafter simply "Condition"), as well as all subsidiary two-way interactions and main effects. Conformity to the harmony phonotactics was coded as the reference level for the categorical predictors. Modeling began by fitting a maximally-specified model (following Barr et al. (2013)) and simplifying as necessary to achieve convergence. The final model contained the fixed effects outlined above, plus random intercepts for participant and nonword.

5.4. Results

The results of the experiment are plotted in Figure 2. Statistical analysis revealed that violating the nasal-harmony phonotactic was associated with significantly lower ratings ($\beta = -24.93$, $p < 0.001$). The interaction between violation of nasal-harmony and Condition was significant ($\beta = 0.29$, $p < 0.001$), indicating that as the percentage of forms violating the nasal-harmony phonotactic in the training data increased, novel forms which violated this phonotactic were judged less ill-formed. The analogous main effects and interaction between violation of the backness-harmony phonotactic and training group was also significant (main effect: $\beta = -9.95$, $p = 0.015$; interaction: $\beta = 0.19$, $p < 0.001$). There was also a significant main effect of training group, indicating that as as the number of fully-conforming words heard in training decreased, fully-conforming words were judged less well-formed as a baseline ($\beta = -0.18$, $p < 0.001$).

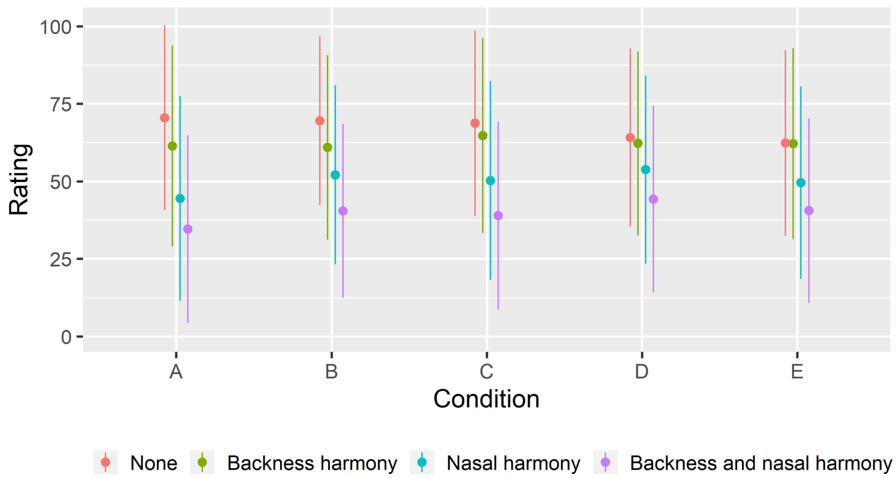

Figure 2: Experiment 1 results, group-level rating plotted on the vertical axis with standard deviation, Condition plotted on the horizontal axis. Color denotes which phonotactics were violated.

Critically, the three-way interaction between violation of nasal-harmony, violation of backness-harmony, and training group was significant ($\beta = -0.17$, $p < 0.002$). As the percentage of nonconforming words in training increased, the difference between singly-marked and unmarked items *decreased*, while the relative markedness associated with the doubly-marked items remained approximately unchanged.

For the sake of comparison, we can visualize what the doubly-violating data would look like if they were subject to linear cumulativity of penalties from the individually-violating forms. We re-fit a mixed-effects model with main effects of violating backness and nasal harmony, as well as Condition, and the interactions between individual violation and Condition. We then used this model to predict expected ratings for forms which violated *both* backness and nasal harmony (the magenta points in figure 3) using `lme4`'s *predict()* function (Bates et al., 2015b).[3] The divergence between this series and the the attested rating for doubly-violating forms (the teal points in figure 3) that increases with the number of exceptions in the training data can be interpreted as the degree of

[3]We predicted ratings based on group-level predictors only (corresponding setting: re.form = NULL), and did not restrict the model to predicting values only for previously-observed violation profiles (allow.new.levels = TRUE).

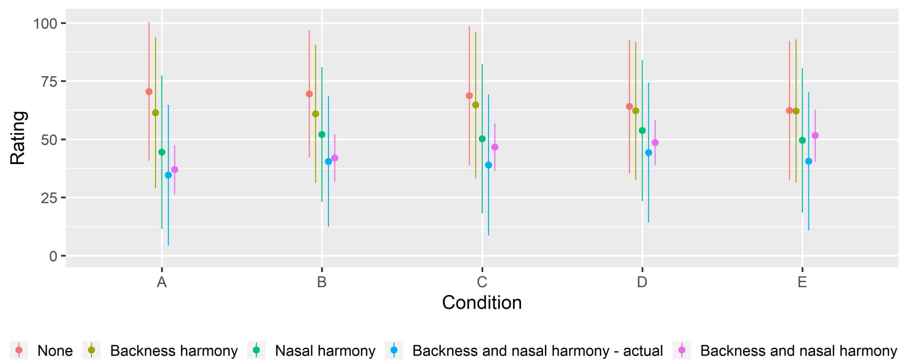

Figure 3: Experiment 1 results with predicted linear effect (magenta) compared to attested super-linear effect (teal).

super-linear cumulativity displayed.

5.5. Local discussion

Experiment 1 tested for a causal link between the strength of a given phonotactic in isolation and the linearity of cumulative constraint interactions in which it participates. We found that learners judged doubly-violating items as more ill-formed than expected when trained on a language which contained more exceptions to individual phonotactic principles than when trained on a more categorical language, supporting such a connection. The findings mirror the phonotactic typology outlined in Albright (2012), where words that violate two weakly-enforced phonotactic generalizations are less frequent than would be expected by the joint probability of those two constraint violations. Because the relationship between phonotactic strength and super-linear cumulativity emerged in a poverty-of-the-stimulus paradigm, we conclude that the similarities between these experimental findings and the underattestation of such items in natural language lexicons are unlikely to be due to chance or a confounding third factor, and instead have their source in a synchronic grammar which is characterized by a non-linear relationship between Harmony and probability.

5.6. Against an under-learning explanation

Because the fully-conforming forms became indistinguishable in generalization from those violating only the backness-harmony phonotactic at the highest level of exceptionality (Condition E), it is possible that participants simply *did not* *learn* the backness harmony phonotactic. We deem this unlikely, because participants were required to learn both phonotactics to non-significantly-different degrees, per the verification phase. Despite this, it may be that a preponderance of participants in this Condition were simply skating through at chance on one phonotactic and above chance on the other.

To test for this possibility, we calculated each participant's *nasal advantage* score, a measure ranging between -2 and 2 which corresponded to the difference between the number of correct answers (out of 8) that participant gave on questions testing backness- vs. nasal-harmony in the verification phase. A positive score indicates that a participant got more correct answers on the nasal-harmony-assessing questions, and a negative answer indicates the reverse. If participants were simply not learning the backness-harmony phonotactic, we should expect to see participants in training Conditions wither more exceptions having a higher *nasal advantage* score. Figure 4 plots nasal advantage scores by Condition.

A linear model confirmed the visual impression that training Condition (coded as a numerical predictor corresponding to the percentage of training data conforming to both phonotactics) does not significantly predict nasal advantage score ($\beta = -0.015$, $p = 0.791$). We therefore deem it unlikely that the effects observed in Experiment 1 are due to *insufficient learning* of the backness harmony phonotactic.

5.7. Was super-linear cumulativity emergent or overtly learned?

Recall that as the number of singly-violating forms in training increased across Conditions, the number of doubly-violating forms in training remained zero. This design maintained a poverty-of-the-stimulus environment, ensuring that generalization to doubly-violating forms was unaffected by training data. A side-effect of this choice, however, was that the training data contained subtle super-linear under-representation of doubly-marked forms, raising the possibility that the super-linear cumulativity observed in participants' responses was explicitly learned from the data. Put another way, the more individual exceptions to each phonotactic a participant sees, the more conspicuous the absence of a doubly-violating form becomes. In the most exceptionful training phase, Condition E, participants would expect to hear 2 doubly-violating tokens of the form *poni* based on the joint probability of backness-harmony-violating and nasal-harmony-violating stimuli in their training data. Instead they heard zero — a blatant case of super-linear underrepresentation of doubly-marked forms, exactly of the type discussed in section 3.1.1. Thus the results of Experiment 1

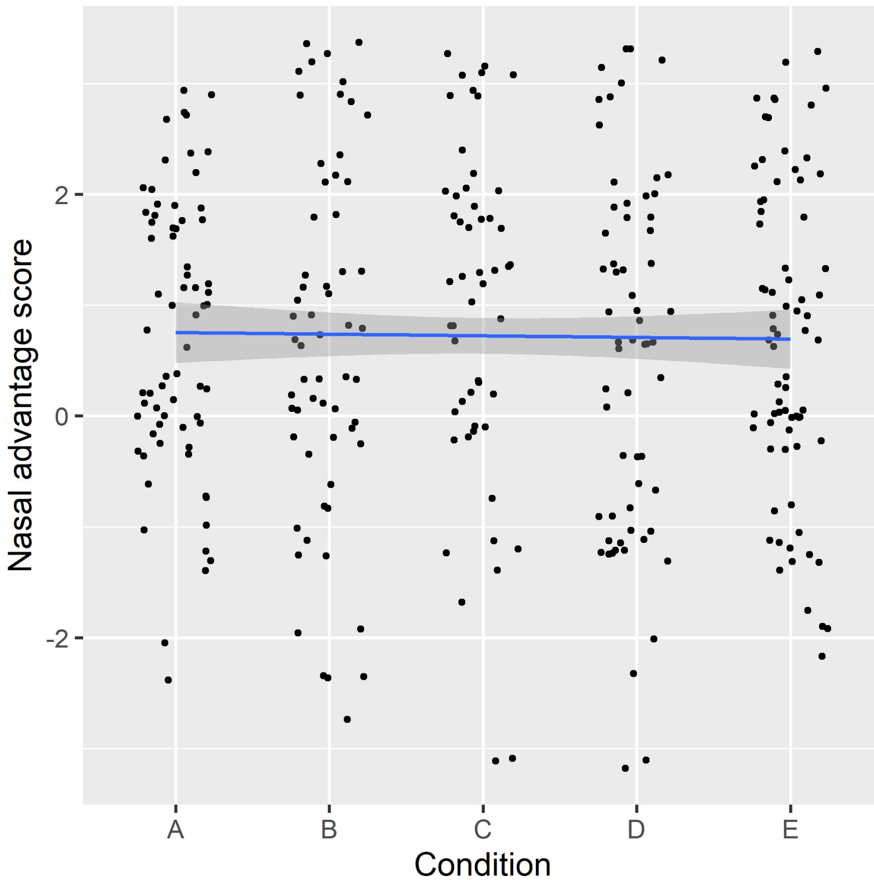

Figure 4: Nasal advantage score by Condition: one dot is one participants' score (jitter added for readability).

are ambiguous between supporting a synchronic grammar which incorporates a non-linearity between Harmony and probability that allows super-linear cumulativity to "emerge" under certain weighting conditions, or simply indicating that learners are sensitive enough to infer a conjoined constraint on the basis of (maximally) 2 missing forms in their learning data (cf. Shih (2017) for more on the role of constraint conjunction in yielding non-linearity in weighted-constraint grammars). We conducted a follow-up experiment to distinguish between these two possibilities.

	Experiment:	1E	2
Unmarked	*potu*	16	16
Back exceptions	*poti*	8	8
Nasal exceptions	*ponu*	8	8
Doubly-violating	*poni*	0	2

Table 2: Distribution of training items by type, comparing Experiment 1 (Condition E) to Experiment 2.

6. Experiment 2: a control for Condition E

To determine whether the super-linear cumulativity observed in Experiment 1 was emergent or overtly learned, we carried out a replication of Condition E from Experiment 1, except that the training data included two doubly-violating forms, thus removing their super-linear underrepresentation in training data. If participants in Experiment 2 exhibit constraint linearity which significantly differs from that exhibited in Experiment 1 Condition E, we can conclude that the source of the effect must lie in the structure of the phonological grammar, which would support the synchronic view outlined in 3.1.1. On the other hand, if participants exhibit linear cumulativity in this experiment, we can conclude that the participant is sensitive enough to detect such patterns in the learning data, suggesting support for the diachronic view from section 3.1.1, paired with a highly-attuned learning mechanism.[4]

6.1. Methods

6.1.1. Participants

86 undergraduate students were recruited to participate in the experiment, of which 15 were excluded for not being native speakers of English, leaving data from 71 participants for analysis. The current experiment only had one training Condition which was identical to that used in Experiment 1's Condition E, except that two of the singly-violating forms were altered so as to also violate the other phonotactic; see table 2. Compensation, design, and procedure were identical to those of Experiment 1.

[4]Note that even if speakers do not distinguish between Experiment 1's Condition E and Experiment 2, it may still be possible for speakers to overtly learn a super-linear pattern in their data; all we will have demonstrated is that *in this case* the super-linear cumulativity in Experiment 1 was not due to overt learning.

6.2. Analysis

To test whether the linearity of cumulativity exhibited in Experiment 2 differed meaningfully from of Condition E of Experiment 1, the two datasets were analyzed together in a mixed-effects linear regression model. In contrast to Experiment 1, we used a Bayesian implementation of the model so as to allow for direct interpretabilty of a possible null result, using the `brms` package (Bürkner et al., 2017).[5] Bayesian models estimate a range of probable values for the parameters of interest directly; thus we can conclude that an effect is robust to the extent that 95% of these values, a measure known as a 95% Credible Interval (abbreviated to "95% CI", followed by upper and lower bounds in brackets), does not include zero. The inverse of this is that if this range is centered on zero, then we can say there is no evidence of an effect for the parameter of interest. For a linguistically-oriented introduction to Bayesian methods for both theory-building and data analysis, see Nicenboim and Vasishth (2016); for tutorial materials on the `brms` package in a linguistic context, see Nalborczyk et al. (2019); Vasishth et al. (2018); for a more general primer in Bayesian modeling, see Kruschke (2014).

As in Experiment 1, the dependent variable was the numerical rating given to each word in the generalization phase. Also as in Experiment 1 the model contained a fixed effect of whether the form violated backness harmony (*y*/*n*), whether the form violated nasal harmony (*y*/*n*), and a binary factor for Experiment (*one*/*two*), as well as all two- and three-way interactions of these predictors. The model also contained random intercepts for word with slopes for Experiment, and random intercepts for subject with slopes for the interaction of the two binary phonotactic factors.

We can interpret the results of the model as follows: if the 95% Credible Interval for the three-way interaction of violating backness harmony, violating nasal harmony, and Experiment excludes zero, it indicates that the degree of linearity in the cumulative interaction of violating both phonotactics together compared to their independent violations differed meaningfully between studies. If the 95% Credible Interval for the interaction includes zero, we can conclude that the cumulative effect of violating both phonotactics did not differ between studies, and thus was unlikely to have been overtly learned in Experiment 1.

[5] The model we fit used default weakly-informative priors, with a burn-in period of 1000 iterations followed by a sampling period of 1000 iterations. We ran four chains to ensure thorough exploration of the posterior distribution, and all $\hat{R}$ values were between 1 and 1.01, indicating that the chains mixed successfully.

6.3. Results

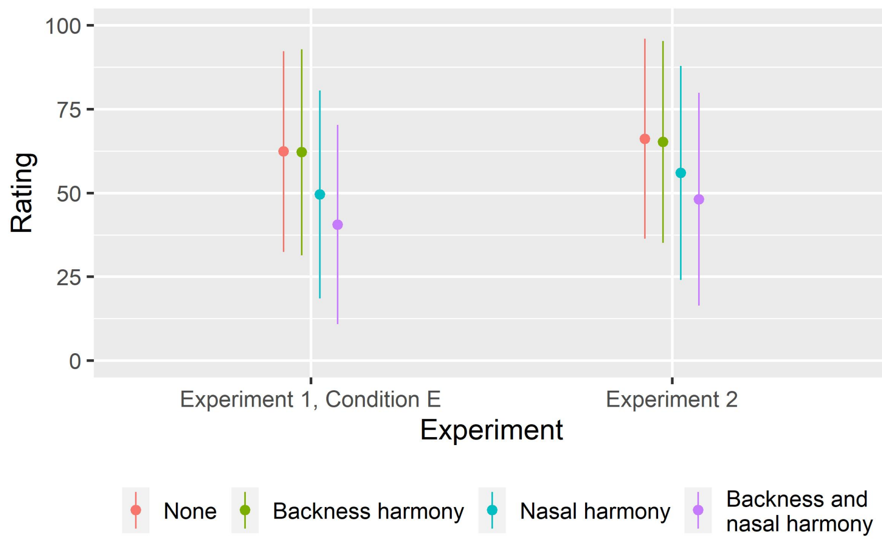

Figure 5: Comparison of mean and standard deviation of ratings by word type in Experiment 1 Condition E, and Experiment 2.

The results of Experiment 2 are shown in Figure 5. Violating nasal harmony resulted in lower ratings ($\beta = -12.72$, 95% CI $[-21.49, -3.85]$), while the effect of violating backness harmony was not ($\beta = -0.10$, 95% CI $[-8.91, 8.72]$). One experiment was not reliably associated with higher ratings than the other overall ($\beta = 3.89$, 95% CI $[-0.87, 8.35]$). The coefficient for the interaction between violating backness and nasal harmony did not differ meaningfully from zero ($\beta = -8.87$, 95% CI $[-22.44, 4.96]$), nor did the coefficient for the interaction between Experiment and violating backness harmony ($\beta = -0.72$, 95% CI $[-4.50, 2.96]$), nor did the coefficient for the interaction between Experiment and violating nasal harmony ($\beta = 2.56$, 95% CI $[-2.30, 7.22]$). Turning to the quantity of interest, the credible intervals for coefficient of the three-way interaction between violating backness harmony, violating nasal harmony, and Experiment centered around zero ($\beta = 1.89$, 95% CI $[-4.13, 7.73]$).

6.4. Local discussion

Experiment 2 tested for whether the super-linear cumulativity observed in Experiment 1 was a result of participants overtly learning the super-linear cumulativity in their data. We found that the linearity of cumulativity was not affected by whether or not the training data used contained a subtle super-linear pattern, a possible confound to the results of Experiment 1. We take this to be compelling evidence in support of a synchronic link between exceptionality in learning data and super-linear cumulativity, as discussed in Section 3.1.1, and against the possibility of the effect having been overtly learned. With this data in hand, we turn to how these data can be accounted for by phonological theories.

7. Incorporating super-linearity into phonological models

Since we observed that super-linear cumulativity emerges from exceptionful learning data in a poverty-of-the-stimulus paradigm, we can ask what qualities a phonological theory must possess in order to capture this relationship. In section 3.1.2 we noted that MaxEnt and NHG frameworks have been shown to be able to relate Harmony and probability in a sigmoidal fashion; in the remainder of the paper we assess these frameworks in light of the experimental data. We also demonstrate that other frameworks which do not display this non-linear relationship fare substantially worse in capturing the data.

7.1. Two questions to ask when modeling experimental data

Phonological theories can be evaluated on their ability to model experimental results in (at least) two ways: one is to ask whether, given the experimental results, a model can achieve a good fit to the experimental data. Success on this metric indicates that the empirical data are in the space of grammars which are allowed under the framework in question — that is, the framework in question is *descriptively adequate* with respect to the phenomenon, and the model needs no further aid from linking hypotheses to capture the data. Given that a framework *can* capture the experimental results, another more rigorous criterion is *explanatory adequacy*: the proposed model should reproduce the crucial data pattern in the experimental findings given the same training data as the experimental subjects. In a sense, this is asking the model to "take the experiment" itself. Below, we evaluate MaxEnt and NHG using these criteria.

7.1.1. Model setup and fitting

We adopt a linking hypothesis which interprets the experimental rating (0-100) as the probability of endorsement in a binary lexical decision task (0%-100%). Therefore, we employ a model structure in which each attested form competes with a single alternative candidate, the Null Parse (Prince and Smolensky, 1993; Albright, 2008, 2012). An example of one such competition is in table 3.

/poti/	AGR([back]) 2.5	MPARSE 1	H
[poti]	1		2.5
☞*null*		1	1

Table 3: Competition with the null parse in a schematic tableau.

Our choice to model the experimental data as competition with the Null Parse is motivated the fact that it is in competition with another candidate that MaxEnt and NHG exhibit a sigmoidal relationship between probability and acceptability, and thus we give the models the best chance to succeed. Although often used in cases of phonologically-conditioned gaps in morphological or syntactic contexts (Rice, 2005; Prince and Smolensky, 1993; McCarthy, 2005, *among others*), in the context of phonotactics the Null Parse model can be thought of as setting a minimum threshold of well-formedness as a criterion of existence. In the context of a probabilistic model of phonotactics, this threshold is "soft", allowing exceptions but dramatically reducing likelihood of forms with a Harmony penalty greater than the weight of MPARSE. Despite a superficial similarity to traditional faithfulness-based approaches to phonotactics (Prince and Smolensky, 1993; Tesar and Smolensky, 1998; Prince and Tesar, 2004), in the Null Parse model the phonotactic grammar allots a probability to a possible form, the alternative to which is nonexistence (rather than repair), as in the markedness-only approach to phonotactics taken by Hayes and Wilson (2008); Wilson and Gallagher (2018) and others.

7.2. Evaluating descriptive adequacy

We first evaluated the descriptive adequacy of MaxEnt and NHG to capture the results of Experiment 1. Figure 7.2 lists predicted ratings derived from the statistical model in Section 5.4.[6]

For MaxEnt and NHG, we fit three parameters for each Condition: the weights of MPARSE, AGREE([nas]), and AGREE([back]).[7] Figure 7.2 displays the experi-

[6]We used model-predicted group means per Condition instead of simply averaging over ratings for each category so as to normal potential by-participant and by-item idiosyncrasies in the experimental results.

[7]The MaxEnt models were fit using the Solver utility in Microsoft Excel (Fylstra et al., 1998). The NHG models were fit using OTSoft (Hayes et al., 2003), with the following settings: 1,000,000 learning iterations, 100,000 testing iterations, initial plasticity = 0.01, final plasticity = 0.001;

mental findings alongside the best-fitting predictions of the MaxEnt and NHG models; these values are also given in table 7.2.

		Condition				
Source	*Phonotatics violated*	A	B	C	D	E
Experiment 1	None	0.70	0.67	0.68	0.64	0.62
	Nasal	0.44	0.52	0.50	0.54	0.49
	Backness	0.61	0.61	0.64	0.62	0.62
	Backness and nasal	0.35	0.40	0.38	0.44	0.40
MaxEnt	None	0.70	0.70	0.70	0.66	0.64
	Nasal	0.44	0.51	0.48	0.52	0.47
	Backness	0.61	0.60	0.62	0.60	0.60
	Backness and nasal	0.35	0.41	0.40	0.46	0.42
NHG	None	0.77	0.78	0.79	0.80	0.76
	Nasal	0.43	0.53	0.55	0.55	0.51
	Backness	0.51	0.53	0.55	0.55	0.51
	Backness and nasal	0.29	0.36	0.38	0.39	0.35

Table 4: Mean group-level percent endorse for Experiment 1 (converted from predicted ratings by dividing by 100, obtained from the regression model in section 5), and best-fitting predicted values from MaxEnt and NHG grammars.

The two sets of experimental findings and model predictions were then compared by calculating the sum of the squared error between the predicted and observed data; these are given on table 5.

Model	*Sum squared error*	*Mean squared error*	*Max squared error*
MaxEnt	0.0057	0.0003	0.0008
NHG	0.1264	0.0063	0.0240

Table 5: Sum, mean, and maximum squared error for the MaxEnt and NHG models.

noise applied by-constraint.

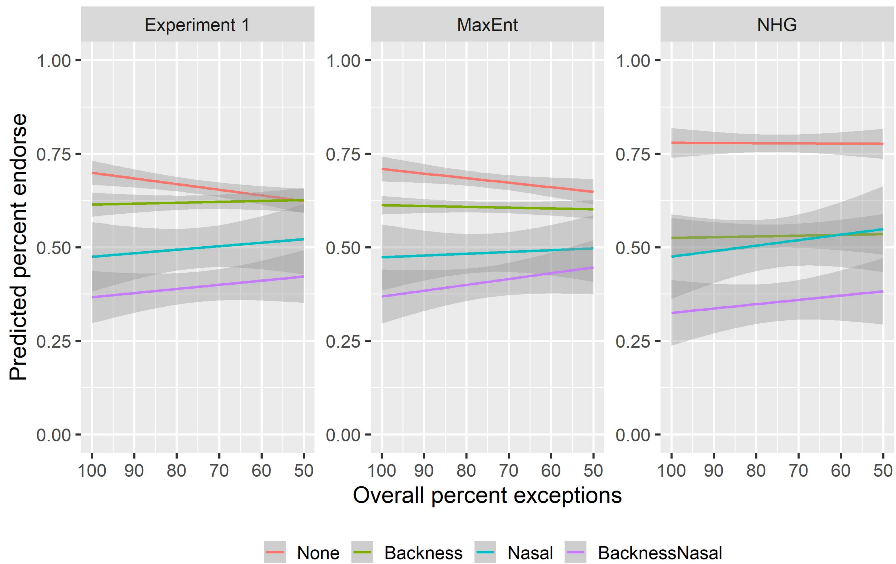

Figure 6: Attested (left) and model predicted (MaxEnt, center; NHG, right) experimental results. For ease of comparison with the MaxEnt and NHG predictions, model-predicted mean values per stimulus type are plotted across a continuous horizontal axis corresponding to Condition for the Observed data.

It is clear from both Figure and table 5 that while the MaxEnt model achieves a qualitatively and quantitatively tight fit to the experimental data, the NHG matches its training data with less fidelity. Why should this be? As noted in Smith and Pater (2020), although NHG is able to capture super-linear cumulativity, its ability to do so is more limited than MaxEnt. We demonstrate in the next section, however, that both the MaxEnt and NHG models fare significantly better than other models which do not possess a non-linear relationship between Harmony and probability.

7.3. Evaluating other models

To this point we have considered only those frameworks which are known to have a sigmoidal relationship between Harmony and probability — MaxEnt and NHG — as candidate models of the phonological grammar. There remain, however, a number of other phonological frameworks commonly in use which have not been considered, Stochastic OT (Boersma et al., 1997, 1998; Boersma and Hayes, 2001), Linear Harmonic Grammar (Coetzee and Pater (2008), cf. also the nearly-identical Linear Optimality Theory of Keller (2006)), and the surface-based

Maximum Entropy approach (Hayes and Wilson, 2008; Wilson and Gallagher, 2018). So as to avoid an abundance of theories being underdetermined by a dearth of discriminating data, we briefly take up these models here, despite having no *a priori* reason to believe they should perform particularly well, since they do not exhibit the non-linear relationship between Harmony and probability that MaxEnt and NHG do. We demonstrate that these models fare substantially worse on the measure of descriptive adequacy, using the same linking hypothesis as in section 7.1.1.

7.3.1. Model setup and fitting

The Stochastic OT model was fit using OTSoft with the same settings as the NHG model. Fitting the Linear HG model followed the protocol for the MaxEnt model described above, except that the negative Harmony of each candidate was not exponentiated in the calculation of the probability of occurrence. This reduces each binary competition to an instance of the Luce Choice Axiom (Luce, 1959), as previously noted by Pizzo (2015). The markedness-only model necessitated a slightly more involved approach: because the model yields a probability distribution over candidates which must sum to 1, each of the five separate MaxEnt models without the Null Parse candidate contained four candidates, corresponding to a fully-conforming stimulus (*potu*), two singly-violating stimuli (*poti, ponu*) and a doubly-violating stimulus (*poni*). Constraints and violations were assessed as in the models discussed in section 7.1.1, with the exception that the markedness-only models contained no MParse constraint because the null parse was not a candidate. Finally, to allow direct comparison between the model output and Experiment 1, the probability assigned to each word type was raised to the power of 1/T, where T is a free *temperature* parameter.[8] The models were fit in Excel using Solver, with the objective function of maximizing the sum of the log-likelihood of the data (by allowing the constraint weights to vary, as is standard in MaxEnt) plus the negative sum of the squared error between the model's predicted probabilities for each category and those obtained in Experiment 1 (with T free to vary to allow for a best-fit scaling parameter for each Condition's model).

These procedures yielded the following fits, plotted in Figure 7 and displayed in table 6.

[8]For prior usage of *temperature* as a scaling factor in comparing model-generated and experimentally-obtained data, see Hayes and Wilson (2008), Breiss (*submitted*).

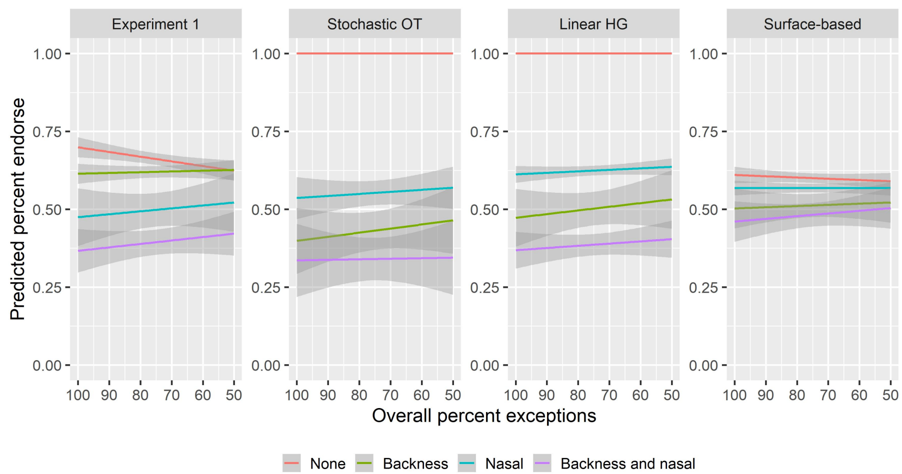

Figure 7: Attested (left) and model predicted (Stochastic OT, center left; Linear HG, center right; Surface-based model, right) experimental results. As in Figure 7.2, model-predicted mean values per stimulus type are plotted across a continuous horizontal axis corresponding to Condition for the Observed data.

Summary statistics for these models, as well as the best-performing MaxEnt model from table 5, is shown in table 7.

In line with the discussion in section 3.1.2, all of the models which do not explicitly allow for a non-linear relationship between Harmony and acceptability fall substantially short of even the less-well-fitting NHG model.

7.4. Evaluating explanatory adequacy

Since we have observed that the experimental data are approximated well by the null-parse MaxEnt model, we can ask whether the relationship between exceptionality in learning data and super-linear cumulativity emerges from the framework given the same training data as the human participants, as suggested in section 3.1.2.[9] We fit a set of 5 MaxEnt grammars, providing each with the

[9]Though in principle this same test could be carried out with NHG models, it's not clear what "success" would look like, given the apparent remoteness of the experimental results from the framework's parameter space.

		Condition				
Source	*Phonotactics violated*	A	B	C	D	E
Experiment 1	None	0.70	0.67	0.68	0.64	0.62
(for reference)	Backness	0.61	0.61	0.64	0.62	0.62
	Nasal	0.44	0.52	0.50	0.54	0.49
	Backness and nasal	0.35	0.40	0.38	0.44	0.40
Stochastic OT	None	1	1	1	1	1
	Backness	0.55	0.51	0.56	0.58	0.55
	Nasal	0.39	0.41	0.42	0.51	0.42
	Backness and nasal	0.30	0.39	0.32	0.38	0.32
Linear HG	None	1	1	1	1	1
	Backness	0.61	0.61	0.64	0.63	0.63
	Nasal	0.44	0.52	0.50	0.55	0.50
	Backness and nasal	0.35	0.39	0.39	0.42	0.38
Surface-based	None	0.60	0.61	0.61	0.60	0.58
	Backness	0.56	0.57	0.58	0.57	0.56
	Nasal	0.48	0.53	0.51	0.54	0.50
	Backness and nasal	0.44	0.49	0.48	0.52	0.48

Table 6: Mean group-level percent endorse for Experiment 1 (converted from predicted ratings by dividing by 100, obtained from the regression model in section 5), and best-fitting predicted values from Stochastic OT, Linear HG, and Surface-based grammars.

Model	*Sum squared error*	*Mean squared error*	*Max squared error*
MaxEnt (for reference)	0.0057	0.0003	0.0008
Stochastic OT	0.6438	0.0322	0.1444
Linear HG	0.5638	0.0212	0.1444
Surface-based	0.0765	0.0038	0.0105

Table 7: Sum, mean, and maximum squared error for the Stochastic OT, Linear HG, and Surface-based models; the best-fitting MaxEnt model is repeated from table 5 for reference.

training data of one of the Conditions in Experiment 1 (as in table 1). Each grammar was allowed to set the weights of MParse, Agree([nas]), and Agree([back]) to match the frequencies in training data as closely as possible — this corresponds to maximizing the likelihood of the training data, as is standard with MaxEnt models. However, because participants exhibited markedly different responses to violations of AGREE([nas]) and AGREE([back]), we additionally require the model to match the average difference in ratings between nasal-harmony and backness-harmony violators across Conditions, as measured using the sum standard error, by modifying the likelihood term in the model. Figures 8 and 9 display these results.

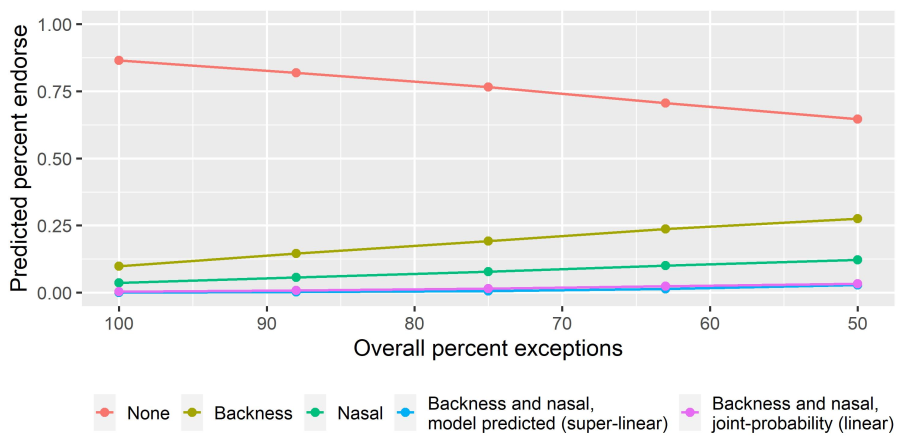

Figure 8: MaxEnt predictions based on Experiment 1's training data.

To demonstrate that super-linear cumulativity emerges from the MaxEnt model as discussed in section 3.1.1, we plot the predicted percent endorsement rate of the doubly-violating forms (purple) based on the joint probability of their individual violations — that is, the percent endorse based on linear cumulativity — alongside the actual predicted endorsement rate under the model (blue).

8. General discussion

This paper sought to determine whether super-linear cumulativity observed in prior experimental work had its source in the synchronic grammar, or whether the data patterns must be explained via frequency matching a diachronically-influenced lexicon. Examining the predictions of two constraint-based phonological frameworks, MaxEnt and NHG, we found that under certain conditions these

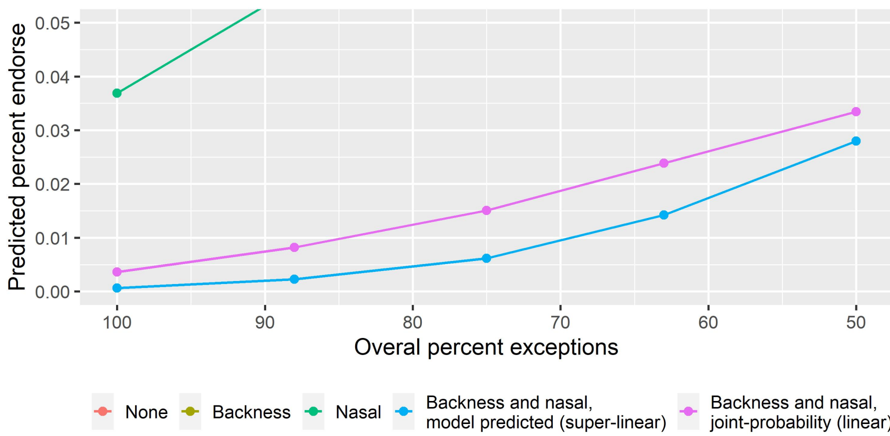

Figure 9: Figure 8, zoomed to focus on the doubly-violating forms, demonstrating emergent super-linear cumulativity in the MaxEnt model.

frameworks exhibit a super-linear relationship between Harmony and probability. Further, these models predict that any synchronic super-linearity of the constraint interaction should be sensitive to the strength of the constraints involved, operationalized here as relative exceptionfulness. We tested this prediction using an AGL experiment in which we systematically varied the number of exceptions to the two primary harmony phonotactics that learners were exposed to in training. We found that as the number of exceptions to each phonotactic in training increased, the well-formedness of the doubly-violating forms became lower than predicted based on the linear sum of their penalties — that is, participants exhibited a super-linear cumulativity of markedness violations, modulated by the strength of the constraints involved. Because the effect emerged in a poverty-of-the-stimulus paradigm using an artificially-controlled language, we concluded that we must attribute the finding to the synchronic grammar, rather than simply to a side-effect of a diachronically-skewed lexicon. We also verified that this super-linear behavior was *emergent* from the interaction of the two constraints — a property of the grammar itself — rather than overtly learned from the training data.

Building on these experimental findings that implicated the synchronic grammar, we evaluated a range of probabilistic constraint-based phonological models on their ability to capture the data. We found that while a MaxEnt model was able to capture the participants' responses well, the NHG model fell short. Further, as predicted, super-linear cumulativity emerged from the MaxEnt framework when exposed to the same exceptionful training data as the experimental participants.

8.1. The mathematical basis of super-linearity in MaxEnt

Because the MaxEnt framework is based on relatively simple equations relating Harmony and probability, we can explicitly outline the weighting conditions under which a MaxEnt grammar will depart from linear cumulativity. To start, we use variables to abstract away from specific constraint weights, and thus can characterize the probability of the singly-marked *poti*, violating AGR([back]), as being proportional to $e^{-w(\textsc{Agr}([back]))}$, and that of the other singly-marked type *ponu*, violating AGR([nas]), as being proportional to $e^{-w(\textsc{Agr}([nas]))}$. By the same logic, the probability of the doubly-marked type *poni*, violating both AGR([back]) and AGR([nas]), is proportional to $e^{-w(\textsc{Agr}([back]))} + e^{-w(\textsc{Agr}([nas]))}$, and the probability of the Null Parse is proportional to $e^{-w(\text{MP}\textsc{arse})}$.

It follows, therefore, that the probability of the backness-violating *poti*, when competing against the Null Parse, should be $\frac{e^{-w(\textsc{Agr}([back]))}}{Z}$, where Z is $e^{-w(\textsc{Agr}([back]))} + e^{-w(\text{MP}\textsc{arse})}$, and that the probability of nasal-violating *ponu* in its own competition against the Null Parse is $\frac{e^{-w(\textsc{Agr}([nas]))}}{Z}$, where Z is $e^{-w(\textsc{Agr}([nas]))} + e^{-w(\text{MP}\textsc{arse})}$. Continuing in this vein, the probability of the doubly-violating form *poni* in competition with the Null Parse is $\frac{e^{-w(\textsc{Agr}([nas]))-w(\textsc{Agr}([back]))}}{Z}$, where Z is $e^{-w(\text{MP}\textsc{arse})} + e^{-w(\textsc{Agr}([back]))-w(\textsc{Agr}([nas]))}$. In contrast, we can can obtain the *joint probability* of violating both constraints by multiplying the probability of the forms with each of those individual violations. This is shown in equation 1.

$$\frac{e^{-w(\textsc{Agr}([nas]))}}{e^{-w(\text{MP}\textsc{arse})} + e^{-w(\textsc{Agr}([nas]))}} \times \frac{e^{-w(\textsc{Agr}([back]))}}{e^{-w(\text{MP}\textsc{arse})} + e^{-w(\textsc{Agr}([back]))}} \tag{1}$$

This expression simplifies to the following:

$$\frac{e^{-w(\textsc{Agr}([back]))-w(\textsc{Agr}([nas]))}}{\left(e^{-w(\text{MP}\textsc{arse})} + e^{-w(\textsc{Agr}([back]))}\right)\left(e^{-w(\text{MP}\textsc{arse})} + e^{-w(\textsc{Agr}([nas]))}\right)} \tag{2}$$

This equation simplifies again, and allows us to characterize the joint probability of two markedness violations as the following:

$$\frac{e^{-w(\textsc{Agr}([back]))-w(\textsc{Agr}([nas]))}}{\left(e^{-w(\text{MP}\textsc{arse})-w(\textsc{Agr}([back]))} + e^{-w(\text{MP}\textsc{arse})-w(\textsc{Agr}([nas]))} + e^{-w(\textsc{Agr}([nas]))-w(\textsc{Agr}([back]))} + e^{-2w(\text{MP}\textsc{arse})}\right)} \tag{3}$$

Comparing this quantity to the probability of the doubly-marked candidate in its own competition against the Null Parse, it becomes clear why certain weighting conditions in MaxEnt yields super-linear cumulativity: the denominators in equations 3 and 4 are not the same.

$$\frac{e^{-w(\textsc{Agr}([back]))-w(\textsc{Agr}([nas]))}}{e^{-w(\text{MP}\textsc{arse})}+e^{-w(\textsc{Agr}([back]))-w(\textsc{Agr}([nas]))}} \tag{4}$$

Because Harmony is computed via the simple addition of penalties before exponentiation in equation 4, the probability of the doubly-violating candidate *poni* is not guaranteed to equal the joint probability of the structures which make it marked (*poti, ponu*). We can examine the relationship between these two quantities by cancelling the $e^{-w(\textsc{Agr}([back]))-w(\textsc{Agr}([nas]))}$ term found in the numerator and denominator out of both equations 3 and 4, and compute the ratio of the remaining quantities to see exactly when MaxEnt will exhibit super-linear cumulativity of markedness violations.

$$\frac{e^{-w(\text{MP}\textsc{arse})}}{e^{-w(\text{MP}\textsc{arse})-w(\textsc{Agr}([back]))}+e^{-w(\text{MP}\textsc{arse})-w(\textsc{Agr}([nas]))}+e^{-2w(\text{MP}\textsc{arse})}} \tag{5}$$

If the ratio in equation 5 is greater than 1, MaxEnt will exhibit super-linear cumulativity of violations: the probability of the doubly-marked candidate will be less than the joint probability of the violating structures in the lexicon. If it is less than 1, MaxEnt predicts sub-linear cumulativity: the probability of the doubly-marked candidate will be greater than the joint probability of the violating structures in the lexicon. Finally, when the ratio is exactly 1 — that is, the denominators in equations 3 and 4 are equal — MaxEnt will exhibit linear cumulativity: the probability of the doubly-marked structure will equal the joint probability of its component structures. Since the weight of MParse is in both quantities in the ratio, the ratio is proportional to the weights of Agr([back]) and Agr([nas]): for any given weight of MParse, relatively higher weights for the two Agree constraints will result in sub-linear cumulativity, and relatively lower weights will result in super-linear cumulativity, as stated in section 3.1.2. Note that, as pointed out in Pizzo (2015, pp. 29-30), a model without exponentiation of Harmony does not exhibit this non-linear relationship - this explains the dramatic underperformance of the otherwise-identical Linear HG model in section 7.1.1.

8.2. Sub-linear cumulativity

To this point we have demonstrated experimentally that there is a non-linear relationship between Harmony and probability. We have also demonstrated that a certain class of phonological theories — MaxEnt and to a lesser extent NHG — exhibit a sigmoidal relationship between Harmony and probability, and thus can capture the data well. We have not demonstrated, however, that the nonlinear relationship between Harmony and probability is necessarily sigmoidal in shape: all that is necessary to account for the experimental data is for the relationship between Harmony and probability to be some nonlinear function with a "shoulder" in it, resembling the upper half of the sigmoid curve drawn in Figure 1 which gives rise to the super-linear cumulativity of weak markedness violations.

If we take the proposed theoretical mechanism at face value, however, the equations in section 8.1 suggest that will we see not only super-linear cumulativity of weak violations, but also *sub-linear cumulativity* of strong violations. Exactly such a finding comes from Pizzo (2015), who investigates the cumulative effect of violating exceptionless syllable-margin constraints in English, crossing an ill-formed onset with an ill-formed coda. She found that while each violation resulted in a significant decrement in assessed well-formedness, the penalty for a form containing both violations was less than that which is predicted by the summed penalty for each of the markedness violations. Though this finding certainly invites further study, it suggests that the sigmoidal property of MaxEnt or NHG may be not only sufficient but also necessary to describe the properties of the grammar.

8.3. Reconciling grammar and lexicon

In section 3.1.1 we considered the possibility that frequency-matching lexical statistics could explain the experimental findings of super-linearity by Albright (2012) and others. Although the experimental results presented in this paper suggest that this is not the case, the puzzle of super-linear underattestation in the lexicon remains unsolved. Albright's lexical study is not alone: a number of researchers have found evidence that some marked structures exhibit super-linear cumulativity in lexical counts, such that the joint probability of two marked structures is greater than the probability of actual lexical items containing both marked structures. Albright (2008) finds that Lakhota roots which contain multiple structures which are only moderately uncommon, such as consonant clusters and fricatives, co-occur in dramatically fewer roots than predicted by their joint probability. Also in this vein is a study by Yang et al. (2018), who carry out a comparison of English and Mandarin monosyllables and find that the attested sub-lexicons are more well-formed than would be expected by the joint probabilities of their parts. In a slightly different domain, Green and Davis

(2014) find that multiple optional syllable structure simplifications in colloquial Bamana are dramatically less likely to co-occur than the joint-probability of each independent process. Kim (2019), building on Kumagai (2017), demonstrates the cumulative effect of nasals on blocking the inter-morpheme obstruent-voicing process *rendaku* in Japanese compounds which also displays non-linear behavior. Super-linear cumulativity has also been observed in the likelihood of belonging to a specific lexical class (Shih, 2017).

How can we reconcile these lexical findings with the synchronic mechanism demonstrated in this paper? Although it is tempting to extend the synchronic mechanism to account for the lexical data, doing so fails to take into account the diachronic origin of these lexical statistics (cf. Frisch (1996); Beguš (2016)). Although lexical statistics are often advanced as evidence of synchronic phonological knowledge, divergences between lexical statistics and productive grammatical knowledge are well-known (Hayes and White, 2013; Becker et al., 2011, among others). Thus simply observing that a generalization holds of a language's lexicon does not necessarily imply that it enjoys a cognitively real status in the synchronic grammar of its speakers. While there is ample evidence that learners change their language over time (Martin, 2007), it is unlikely that such widespread under-representation can be the effect of a single generation of speakers. We suggest, then, that a full explanation of the lexical findings must take into account the bidirectional relationship between the synchronic phonological grammar and the diachronic trends which shape the lexicon. Such an undertaking is left for future research.

9. Conclusion

The primary contribution of this paper was to advance experimental evidence in favor of a synchronic phonological grammar in which the relationship between Harmony and probability/acceptability is non-linear. We have argued that MaxEnt, and to a lesser extent NHG, can capture the data successfully because they encode a nonlinear — and specifically sigmoidal — relationship between Harmony and probability. This is shown by comparison to a range of other theories which do not predict such a relationship. We outlined the structure and predictions of the best-fitting MaxEnt model, including the conditions in which it does and does not predict super-linear cumulativity of violations. Finally, we adduced evidence of sub-linear cumulativity of strong markedness violations from the literature which further supports the specifically *sigmoidal* shape of the nonlinear relationship between Harmony and probability in MaxEnt and NHG.

The work presented here, however, is only a first step towards a fuller understanding of the empirical and typological landscape of non-linear cumulativity. The proposed sigmoidal shape of the relationship between Harmony and probability makes strong predictions about both the wide scope of constraints that can enter into super-linearly cumulative relationships, and also specific claims about the weighting requirements that must be met for such effects to be observed. A great deal of further empirical research, using natural as well as artificial languages and lexicons, is therefore needed to test these predictions.

References

Albright, A. (2008). Cumulative violations and complexity thresholds.

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

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania.*

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

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

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

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015b). 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*, pages 84–125.

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

Beguš, G. (2016). Unnatural trends in the lexicon: Diachrony and synchrony.

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. et al. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives.* Den HaagHolland Academic Graphics/IFOTT.

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

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

Bürkner, P.-C. et al. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80(1):1–28.

Coetzee, A. W. and Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in muna and arabic. *Natural Language & Linguistic Theory*, 26(2):289–337.

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. (1996). *Similarity and frequency in phonology*. PhD thesis.

Fylstra, D., Lasdon, L., Watson, J., and Waren, A. (1998). Design and use of the microsoft excel solver. *Interfaces*, 28(5):29–55.

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.

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

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

Hayes, B. (2017). Varieties of noisy harmonic grammar. In *Proceedings of the Annual Meetings on Phonology*, volume 4.

Hayes, B., Siptár, P., Zuraw, K., and Londe, Z. (2009). Natural and unnatural constraints in hungarian vowel harmony. *Language*, pages 822–863.

Hayes, B., Tesar, B., and Zuraw, K. (2003). Otsoft 2.1, software package.

Hayes, B. and White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, 44(1):45–75.

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

Hayes, B. P. (2020). Assessing grammatical architectures through their quantitative signatures.

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

Keller, F. (2006). Linear optimality theory as a model of gradience in grammar. *Gradience in grammar: Generative perspectives*, pages 270–287.

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

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

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

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

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, 66(2):81.

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

McCarthy, J. J. (2005). Less than zero: Correspondence and the null output.

McPherson, L. and Hayes, B. (2016). Relating application frequency to morphological structure: the case of tommo so vowel harmony. *Phonology*, 33(1):125–167.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., and Bürkner, P.-C. (2019). An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian. *Journal of Speech, Language, and Hearing Research*, 62(5):1225–1242.

Nicenboim, B. and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—part ii. *Language and Linguistics Compass*, 10(11):591–613.

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

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

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

Rice, C. (2005). Nothing is a phonological fact: Gaps and repairs at the phonology-morphology interface.

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

Smith, B. W. and Pater, J. (2020). French schwa and gradient cumulativity. *Glossa: a journal of general linguistics*, 5(1).

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

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

Tesar, B. and Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, 29(2):229–268.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., and Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of phonetics*, 71:147–161.

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

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

Yang, S., Sanker, C., and Priva, U. C. (2018). The organization of lexicons: A cross-linguistic analysis of monosyllabic words. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 164–173.

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